

# INFINITE HORIZON AVERAGE COST DYNAMIC PROGRAMMING SUBJECT TO TOTAL VARIATION DISTANCE AMBIGUITY

IOANNIS TZORTZIS\*, CHARALAMBOS D. CHARALAMBOUS†, AND  
THEMISTOKLIS CHARALAMBOUS‡

**Abstract.** We analyze the infinite horizon minimax average cost Markov Control Model (MCM), for a class of controlled process conditional distributions, which belong to a ball, with respect to total variation distance metric, centered at a known nominal controlled conditional distribution with radius  $R \in [0, 2]$ , in which the minimization is over the control strategies and the maximization is over conditional distributions. Upon performing the maximization, a dynamic programming equation is obtained which includes, in addition to the standard terms, the oscillator semi-norm of the cost-to-go.

First, the dynamic programming equation is analyzed for finite state and control spaces. We show that if the nominal controlled process distribution is irreducible, then for every stationary Markov control policy the maximizing conditional distribution of the controlled process is also irreducible for  $R \in [0, R_{max}]$ . Second, the generalized dynamic programming is analyzed for Borel spaces. We derive necessary and sufficient conditions for any control strategy to be optimal.

Through our analysis, new dynamic programming equations and new policy iteration algorithms are derived. The main feature of the new policy iteration algorithms (which are applied for finite alphabet spaces) is that the policy evaluation and policy improvement steps are performed by using the maximizing conditional distribution, which is obtained via a water filling solution. Finally, the application of the new dynamic programming equations and the corresponding policy iteration algorithms are shown via illustrative examples.

**Key words.** Stochastic Control, Markov Control Models, Minimax, Dynamic Programming, Average Cost, Infinite Horizon, Total Variational Distance, Policy Iteration

**AMS subject classifications.** 93E20, 90C39, 90C47

**1. Introduction.** The infinite horizon average cost per unit-time discrete-time Markov Control Model (MCM), with deterministic strategies is analysed, in an anthology of papers [1–4, 18]. In such MCMs, the corresponding cost-to-go and the dynamic programming recursions depend on the conditional distribution of the underlying controlled process [5]. This means, any ambiguity of the controlled process conditional distribution will affect the optimality and robustness of the optimal decision strategies.

In this paper, we investigate the effects of any ambiguity in the controlled process conditional distribution on the cost-to-go and dynamic programming. We model the ambiguity in the controlled conditional distributions by a ball with respect to the total variation distance metric, centered at a known nominal controlled conditional distribution with radius  $R \in [0, 2]$ . Then, we re-formulate the infinite horizon average cost MCM using minimax optimization techniques, in which the control strategy seeks to minimize the payoff while the conditional distribution, from the class of total variation distance ball, seeks to maximize it.

We begin our analysis by first considering MCM's defined on finite state and control spaces. By employing certain results from [7], we obtain the characterization of the maximizing conditional distribution and the corresponding dynamic programming equation. The main feature of the maximizing conditional distribution is its characterization via a water-filling solution, which is similar in spirit, to extremum problems encountered in information theory, such as, channel capacity and lossy data compression [8]. This leads to a dynamic

\*Department of Electrical and Computer Engineering, University of Cyprus (UCY), Nicosia, Cyprus. (tzortzis.ioannis@ucy.ac.cy).

†Department of Electrical and Computer Engineering, University of Cyprus (UCY), Nicosia, Cyprus. (chadcha@ucy.ac.cy).

‡Department of Signals and Systems, Chalmers University of Technology, Gothenburg, Sweden. (themistoklis.charalambous@chalmers.se).

programming equation, which includes in its right hand side, the oscillator semi-norm of the cost-to-go or value function, in addition to the standard terms. We show that, if the nominal controlled process distribution is irreducible, then for every stationary Markov control policy the maximizing conditional distribution of the controlled process is also irreducible, the optimal control strategies exists, for  $R \in [0, R_{max}]$ . Moreover, for this range of  $R$ , we derive a new policy iteration algorithm.

Subsequently, we consider general Borel spaces, we invoke a pair of dynamic programming equations (called generalized), and we derive necessary and sufficient conditions of optimality, based on the concept of canonical triplets [9, 12, 16, 17]. This treatment characterizes optimal strategies for any ball of radius  $R \in [0, 2]$ . The main feature of the corresponding policy iteration algorithm (which is applied for finite alphabet spaces), is that the policy evaluation and policy improvement steps are performed using the maximizing conditional distribution.

The remainder of the paper is organized as follows. In Section 1.1, we introduce the classical infinite horizon dynamic programming equation of MCM with an average cost per unit-time optimality criterion, and we briefly discuss the main results derived in the paper. In Section 2, we give some preliminary results concerning the maximization of a linear functional subject to total variation distance. In Section 3, we study the infinite horizon average cost Markov decision problem for finite state and control spaces, and we derive a new dynamic programming recursion and the corresponding policy iteration algorithm. In Section 4, we consider general Borel spaces, and we investigate the infinite horizon average cost Markov decision problem, using the generalized dynamic programming equations. We also introduce a generalized policy iteration algorithm when the state and control spaces are of finite cardinality. In Section 5, we present two examples which illustrate the implications of the the new dynamic programming recursions on the corresponding policy iteration algorithms.

**1.1. Discussion on the Main Results.** In this section, we describe the main results obtained in the paper with respect to the existing literature. Since we treat finite alphabet spaces and Borel spaces, the formulation below is introduced for Borel spaces.

**1.1.1. Dynamic Programming of Infinite-Horizon MCM.** An infinite horizon MCM with deterministic strategies is a five-tuple

$$(1.1) \quad (\mathcal{X}, \mathcal{U}, \{\mathcal{U}(x) : x \in \mathcal{X}\}, \{Q(dz|x, u) : (x, u) \in \mathcal{X} \times \mathcal{U}\}, f)$$

consisting of the following.

- a) **State Space.** A complete separable metric space (called a Polish space)  $\mathcal{X}$ , which models the state space of the controlled random process  $\{x_k \in \mathcal{X} : k \in \mathbb{N}\}$ ,  $\mathbb{N} \triangleq 0, 1, \dots$ .
- b) **Control or Action Space.** A Polish space  $\mathcal{U}$ , which models the control or action set of the control random process  $\{u_k \in \mathcal{U} : k \in \mathbb{N}\}$ .
- c) **Feasible Controls or Actions.** A family  $\{\mathcal{U}(x) : x \in \mathcal{X}\}$  of non-empty measurable subsets  $\mathcal{U}(x)$  of  $\mathcal{U}$ , where  $\mathcal{U}(x)$  denotes the set of feasible controls or actions, when the controlled process is in state  $x \in \mathcal{X}$ , and the feasible state-actions pairs are measurable subsets of  $\mathcal{X} \times \mathcal{U}$ , defined by

$$(1.2) \quad \mathbb{K} \triangleq \{(x, u) : x \in \mathcal{X}, u \in \mathcal{U}(x)\}.$$

- d) **Controlled Process Distribution.** A conditional distribution or stochastic kernel  $Q(dz|x, u)$  on  $\mathcal{X}$  given  $(x, u) \in \mathbb{K} \subseteq \mathcal{X} \times \mathcal{U}$ , which corresponds to the controlled process transition probability distribution.

e) One-Stage-Cost. A non-negative measurable function  $f : \mathbb{X} \mapsto [0, \infty]$ , called the one-stage-cost, such that  $f(x, \cdot)$  does not take the value  $+\infty$  for each  $x \in \mathcal{X}$ .

To ensure the existence of measurable controls we make the following assumption.

ASSUMPTION 1.1. [12]  $\mathbb{K}$  contains the graph of a measurable functions from  $\mathcal{X}$  to  $\mathcal{U}$ ; that is, there is a measurable function  $\varphi : \mathcal{X} \mapsto \mathcal{U}$  such that  $\varphi(x) \in \mathcal{U}(x)$ , for all  $x \in \mathcal{X}$ . The set of all such functions denoted by  $\mathbb{F}$  are called selectors of the multifunction  $x \mapsto \mathcal{U}(x)$ .

We equip the spaces  $\mathcal{X}$  and  $\mathcal{U}$  with the natural  $\sigma$ -algebra  $\mathcal{B}(\mathcal{X})$  and  $\mathcal{B}(\mathcal{U})$ , respectively. For any measurable spaces  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ ,  $(\mathcal{U}, \mathcal{B}(\mathcal{U}))$ , we denote the set of stochastic Kernels on  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$  conditioned on  $(\mathcal{U}, \mathcal{B}(\mathcal{U}))$  by  $\mathcal{Q}(\mathcal{X}|\mathcal{U})$ , and we denote the set of probability distributions on  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$  by  $\mathcal{M}_1(\mathcal{X})$ . Next, we give the definition of deterministic stationary Markov control policies.

DEFINITION 1.2. A deterministic stationary Markov control policy is a measurable function (selector)  $g : \mathcal{X} \mapsto \mathcal{U}$  such that  $g(x_t) \in \mathcal{U}(x_t)$ ,  $\forall x_t \in \mathcal{X}$ ,  $t = 0, 1, \dots$ . The set of such deterministic stationary Markov policies is denoted by  $G_{SM}$ , and the set of all deterministic control policies (i.e., non-stationary, non-Markov) is denoted by  $G$ .

Define the  $n$ -stage expected cost, for a fixed  $x_0 = x$ , by

$$(1.3) \quad J_n^o(g, x) \triangleq \mathbb{E}_x^g \left\{ \sum_{k=0}^{n-1} f(x_k, u_k) \right\}$$

where  $\mathbb{E}_x^g\{\cdot\}$  indicates the dependence of the expectation operation on the policy  $g \in G$  and  $x_0 = x$ . Then, the average cost per unit-time when policy  $g \in G$  is used, given  $x_0 = x$ , is defined by

$$(1.4) \quad J^o(g, x) \triangleq \limsup_{n \rightarrow \infty} \frac{1}{n} J_n^o(g, x).$$

The Markov Control Problem (MCP) is to find a control policy  $g^* \in G$  such that

$$(1.5) \quad J^o(g^*, x) \triangleq \inf_{g \in G} J^o(g, x) = J^{o,*}(x), \quad \forall x \in \mathcal{X}.$$

For finite cardinality spaces  $(\mathcal{X}, \mathcal{U})$ , it is known [12, 14, 16, 20], that if  $f$  is bounded and for all stationary Markov control policies  $g \in G_{SM}$  the transition probability matrix  $Q(z|x, u)$  is irreducible (that is, all stationary policies have at most one recurrent class), then there exists a solution  $V^o : \mathcal{X} \mapsto \mathbb{R}$  and a constant (independent of  $x \in \mathcal{X}$ )  $J^{o,*} \in \mathbb{R}$  such that  $(J^{o,*}, V^o(x))$  is the solution of the dynamic programming (of the infinite-horizon MCP (1.5))

$$(1.6) \quad J^{o,*} + V^o(x) = \inf_{u \in \mathcal{U}(x)} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} Q(z|x, u) V^o(z) \right\}$$

from which existence of optimal policy  $g^* \in G_{SM}$  is obtained. However, if the irreducibility condition is not satisfied (i.e., there is more than one recurrent class), then the dynamic programming equation (1.6) may not be sufficient to give the optimal policy and the minimum cost [12, 16]. In this case, (1.6) is replaced by the following

$$(1.7a) \quad J^{o,*}(x) = \inf_{u \in \mathcal{U}(x)} \left\{ \sum_{z \in \mathcal{X}} Q(z|x, u) J^{o,*}(z) \right\}$$

$$(1.7b) \quad J^{o,*}(x) + V^o(x) = \inf_{u \in \mathcal{U}(x)} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} Q(z|x, u) V^o(z) \right\}.$$

We refer to (1.7a) as the first general dynamic programming equation and to (1.7b) as the second general dynamic programming equation<sup>1</sup>. Note that, the pair of generalized dynamic programming equations (1.7a)-(1.7b) solves the MCP (1.5), without imposing irreducibility of the conditional distribution of the controlled process [12, 16]. Similar results are also known for Borel spaces, by replacing summations in (1.6) and (1.7) by integrals with respect to conditional distribution, while the characterization of the existence of optimal policies is done via canonical triplets [12].

Since the MCP (1.5) and the dynamic programming equation (1.6) are functionals of the conditional distribution of the controlled process, then the optimal strategies  $g \in G$  are obtained based on the assumption of having an accurate knowledge of the conditional distribution  $Q(dz|x, u)$ . Hence, any ambiguity or mismatch of  $Q(dz|x, u)$  from the true conditional distribution will affect the optimality of the strategies.

Motivated by this implication, in this paper we consider the problem discussed in the next section.

**1.1.2. Dynamic Programming of Infinite-Horizon MCM with Total Variation Distance Ambiguity.** Recall the total variation distance between two probability measures,  $\|\cdot\|_{TV} : \mathcal{M}_1(\mathcal{X}) \times \mathcal{M}_1(\mathcal{X}) \mapsto [0, \infty]$ , defined by

$$\|\alpha - \beta\|_{TV} \triangleq \sup_{P \in \mathcal{P}(\mathcal{X})} \sum_{F_i \in P} |\alpha(F_i) - \beta(F_i)|, \quad \alpha, \beta \in \mathcal{M}_1(\mathcal{X})$$

where  $\mathcal{P}(\mathcal{X})$  denotes the collection of all finite partitions of  $\mathcal{X}$ .

In this paper, we will derive the analogues of (1.6) and (1.7a)-(1.7b), for the class of conditional distributions of the controlled process  $Q(dz|x, u)$ ,  $(x, u) \in \mathbb{K}$  which are stationary, and belong to a ball with respect to total variation distance metric, centered at a nominal controlled process distribution  $Q^o(dz|x, u)$ ,  $(x, u) \in \mathbb{K}$ , having radius  $R \in [0, 2]$  (specifically,  $\{Q(dz|x, u) : \|Q(\cdot|x, u) - Q^o(\cdot|x, u)\|_{TV} \leq R\}$ ).

The precise definition is the following.

**DEFINITION 1.3.** For each  $g \in G_{SM}$ , the nominal controlled process  $\{x_t^g : t = 0, 1, \dots\}$  has a stationary conditional distribution defined by

$$\text{Prob}(x_t \in A | x^{t-1}, u^{t-1}) \triangleq Q^o(A | x_{t-1}, u_{t-1}), \quad \forall A \in \mathcal{B}(\mathcal{X}), \quad t = 0, 1, \dots$$

where  $Q^o(\cdot|\cdot, \cdot) \in \mathcal{Q}(\mathcal{X}|\mathbb{K})$ . Given the nominal controlled process and  $R \in [0, 2]$ , the true controlled process conditional distributions are stationary, and belong to the total variation distance ball defined by

(1.8)

$$\mathbf{B}_R(Q^o)(x, u) \triangleq \left\{ Q(\cdot|x, u) \in \mathcal{M}_1(\mathcal{X}) : \|Q(\cdot|x, u) - Q^o(\cdot|x, u)\|_{TV} \leq R \right\}, \quad (x, u) \in \mathbb{K}.$$

Next, we consider the analogue of (1.5). Define the  $n$ -stage expected cost by

$$(1.9) \quad J_n(g, Q, x) \triangleq \mathbb{E}_x^g \left\{ \sum_{k=0}^{n-1} f(x_k, u_k) \right\}$$

and the corresponding maximizing  $n$ -stage expected cost by

$$(1.10) \quad J_n(g, x) \triangleq \sup_{Q(\cdot|x, u) \in \mathbf{B}_R(Q^o)(x, u)} \mathbb{E}_x^g \left\{ \sum_{k=0}^{n-1} f(x_k, u_k) \right\}.$$

<sup>1</sup>Some authors use the term multichain, instead.

Then, the maximizing average cost per unit-time when policy  $g \in G$  is used, given  $x_0 = x$ , is defined by

$$(1.11) \quad J(g, x) \triangleq \limsup_{n \rightarrow \infty} \frac{1}{n} J_n(g, x).$$

The minimax MCP subject to ambiguity defined by (1.8), is to choose a control policy  $g^* \in G$  such that

$$(1.12) \quad J(g^*, x) \triangleq \inf_{g \in G} J(g, x) = J^*(x), \quad \forall x \in \mathcal{X}.$$

A conditional distribution  $Q^*$  that satisfies (1.11) (see also (1.10)) is called a maximizing conditional distribution, a policy  $g^*$  that satisfies (1.12) is called an average cost optimal policy, and the corresponding  $J^*(\cdot)$  is the minimum cost or value function of the minimax MCP.

Next, we introduce an assumption for the minimax MCP defined by (1.12).

ASSUMPTION 1.4.

- (a) The map  $f : \mathcal{X} \times \mathcal{U} \rightarrow \mathbb{R}$  is bounded, continuous and non-negative.
- (b) The set  $\mathcal{U}(x)$  is compact for all  $x \in \mathcal{X}$ .
- (c) The map  $Q^o(A|\cdot, \cdot)$  is continuous on  $\mathbb{K}$  for every Borel set.

Note that it is possible to relax Assumption 1.4, for example,  $f(x, \cdot)$  can be replaced by a lower semi-continuous function on  $\mathcal{U}(x)$  for every  $x \in \mathcal{X}$ , which is non-negative (see [12] for several relaxations).

We derive the following results.

**Dynamic Programming Equations for Finite Alphabet Spaces.** In Section 3, we assume that  $(\mathcal{X}, \mathcal{U})$  are of finite cardinality and we show that if for all stationary Markov control policies  $g \in G_{SM}$ , and for a given total variation parameter  $R \in [0, 2]$ , the maximizing transition probability matrix  $Q^*(g)$  is irreducible, then the dynamic programming equation corresponding to minimax MCP (1.12) is given by

$$(1.13) \quad J^* + V(x) = \min_{u \in \mathcal{U}} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} Q^o(z|x, u) V(z) + \frac{R}{2} \left( \sup_{z \in \mathcal{X}} V(z) - \inf_{z \in \mathcal{X}} V(z) \right) \right\}.$$

The new term entering in the right side of (1.13) is the oscillator semi-norm of the future pay-off.

**Generalized Dynamic Programming Equations for Borel Spaces.** In Section 4, we assume that  $(\mathcal{X}, \mathcal{U})$  are Borel spaces, and we utilize the concept of canonical triplets to establish existence of optimal strategies via the following generalized dynamic programming equations

$$(1.14a) \quad J^*(x) = \inf_{u \in \mathcal{U}(x)} \left\{ \int_{\mathcal{X}} Q^o(dz|x, u) J^*(z) + \frac{R}{2} \left( \sup_{z \in \mathcal{X}} J^*(z) - \inf_{z \in \mathcal{X}} J^*(z) \right) \right\}$$

$$(1.14b) \quad J^*(x) + V(x) = \inf_{u \in \mathcal{U}(x)} \left\{ f(x, u) + \int_{\mathcal{X}} Q^o(dz|x, u) V(z) + \frac{R}{2} \left( \sup_{z \in \mathcal{X}} V(z) - \inf_{z \in \mathcal{X}} V(z) \right) \right\}.$$

Since Borel spaces include finite alphabet spaces, if irreducibility condition is violated, then the existence of optimal strategies is characterized by the finite alphabet version of (1.14a)-(1.14b).

In addition, we obtain the following.

1. Characterize the maximizing conditional distribution corresponding to the supremum in (1.11).
2. Derive new policy iteration algorithms (applied for finite alphabet spaces), in which the policy evaluation and the policy improvement steps are performed by using the maximizing conditional distribution obtained under total variation distance ambiguity constraint.

Finally, in Section 5 we present illustrative examples based on (1.13) and (1.14).

**2. Maximization over Total Variation Distance Ambiguity.** In this section, we recall certain results from [7], concerning the characterization of the extremum problem of maximizing a linear functional subject to total variation distance ambiguity. We use these results to derive the new dynamic programming equations.

Let  $(\mathcal{X}, d_{\mathcal{X}})$  denote a complete, separable metric space (a Polish space), and  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$  the corresponding measurable space, in which  $\mathcal{B}(\mathcal{X})$  is the  $\sigma$ -algebra generated by open sets in  $\mathcal{X}$ . Define the spaces

$$BC(\mathcal{X}) \triangleq \{\text{Bounded continuous functions } \ell : \mathcal{X} \mapsto \mathbb{R} : \|\ell\| \triangleq \sup_{x \in \mathcal{X}} |\ell(x)| < \infty\}$$

$$BC^+(\mathcal{X}) \triangleq \{\ell \in BC(\mathcal{X}) : \ell \geq 0\}.$$

For  $\ell \in BC^+(\mathcal{X})$ , and  $\mu \in \mathcal{M}_1(\mathcal{X})$  fixed, then we have

$$(2.1) \quad L(\nu^*) \triangleq \sup_{\|\nu - \mu\|_{TV} \leq R} \int_{\mathcal{X}} \ell(x) \nu(dx) = \frac{R}{2} \left\{ \sup_{x \in \mathcal{X}} \ell(x) - \inf_{x \in \mathcal{X}} \ell(x) \right\} + \int_{\mathcal{X}} \ell(x) \mu(dx)$$

where  $R \in [0, 2]$ ,  $\nu^*$  satisfies the constraint  $\|\xi^*\|_{TV} = \|\nu^* - \mu\|_{TV} = R$ , it is normalized  $\nu^*(\mathcal{X}) = 1$ , and  $\nu^*(A) \in [0, 1]$  on any  $A \in \mathcal{B}(\mathcal{X})$ . If  $\mathcal{X}$  is a compact set, since  $\ell(\cdot) \in BC^+(\mathcal{X})$  then both the supremum and infimum are attained and they are finite. Define<sup>2</sup>

$$\begin{aligned} x^0 \in \mathcal{X}^0 &\triangleq \left\{ x \in \overline{\mathcal{X}} : \ell(x) = \sup\{\ell(x) : x \in \mathcal{X}\} \equiv \ell_{\max} \right\} \\ x_0 \in \mathcal{X}_0 &\triangleq \left\{ x \in \overline{\mathcal{X}} : \ell(x) = \inf\{\ell(x) : x \in \mathcal{X}\} \equiv \ell_{\min} \right\} \end{aligned}$$

where  $\overline{\mathcal{X}}$  denotes the closure<sup>3</sup> of  $\mathcal{X}$ . Then, the pay-off  $L(\nu^*)$  can be written as

$$(2.2) \quad L(\nu^*) = \int_{\mathcal{X}^0} \ell_{\max} \nu^*(dx) + \int_{\mathcal{X}_0} \ell_{\min} \nu^*(dx) + \int_{\mathcal{X} \setminus \mathcal{X}^0 \cup \mathcal{X}_0} \ell(x) \mu(dx)$$

and the optimal distribution  $\nu^* \in \mathcal{M}_1(\mathcal{X})$ , which satisfies the total variation constraint, is given by

$$(2.3a) \quad \int_{\mathcal{X}^0} \nu^*(dx) = \mu(\mathcal{X}^0) + \frac{R}{2} \in [0, 1]$$

$$(2.3b) \quad \int_{\mathcal{X}_0} \nu^*(dx) = \mu(\mathcal{X}_0) - \frac{R}{2} \in [0, 1]$$

$$(2.3c) \quad \nu^*(A) = \mu(A), \quad \forall A \subseteq \mathcal{X} \setminus \mathcal{X}^0 \cup \mathcal{X}_0.$$

Note that, if  $\mathcal{X}^0$  is empty then  $\nu^*(\mathcal{X}^0) = R/2$  and if  $\mathcal{X}_0$  is empty then  $\nu^*(\mathcal{X}_0) = 0$ .

Next, we elaborate on the form of the maximizing measure for finite and countable alphabet spaces, and its water filling behavior, since we use them to analyze infinite horizon MCP with finite state and control spaces.

<sup>2</sup>We adopt the standard definitions; infimum (supremum) of an empty set to be  $+\infty$  ( $-\infty$ ).

<sup>3</sup>Closure of a set  $\mathcal{X}$  consists of all points in  $\mathcal{X}$  plus the limit points of  $\mathcal{X}$ .

**2.1. The Maximizing Measure for Finite and Countable Alphabet Spaces.** Let  $\mathcal{X}$  be a non-empty denumerable set endowed with the discrete topology. If the cardinality of  $\mathcal{X}$  denoted by  $|\mathcal{X}|$  is finite, then we can identify any  $x \in \mathcal{X}$  by a unit vector in  $\mathbb{R}^{|\mathcal{X}|}$ . Define the set of probability vectors on  $\mathcal{X}$  by

$$(2.4) \quad \mathbb{P}(\mathcal{X}) \triangleq \left\{ p = (p_1, \dots, p_{|\mathcal{X}|}) : p(x) \geq 0, x = 1, \dots, |\mathcal{X}|, \sum_{x \in \mathcal{X}} p(x) = 1 \right\}.$$

That is,  $\mathbb{P}(\mathcal{X})$  is the set of all  $|\mathcal{X}|$ -dimensional vectors which are probability vectors  $\{\nu(x) : x \in \mathcal{X}\} \in \mathbb{P}(\mathcal{X})$ ,  $\{\mu(x) : x \in \mathcal{X}\} \in \mathbb{P}(\mathcal{X})$ . Also, let  $\ell \triangleq \{\ell(x) : x \in \mathcal{X}\} \in \mathbb{R}_+^{|\mathcal{X}|}$  (i.e., the set of non-negative vectors of dimension  $|\mathcal{X}|$ ). Then, (2.1) may be written as follows

$$(2.5) \quad L(\nu^*) = \max_{\nu \in \mathbb{B}_R(\mu)} \sum_{x \in \mathcal{X}} \ell(x) \nu(x)$$

where

$$(2.6) \quad \mathbb{B}_R(\mu) \triangleq \left\{ \nu \in \mathbb{P}(\mathcal{X}) : \|\nu - \mu\|_{TV} \triangleq \sum_{x \in \mathcal{X}} |\nu(x) - \mu(x)| \leq R \right\}.$$

By defining  $\xi(x) \triangleq \nu(x) - \mu(x)$ ,  $x = 1, \dots, |\mathcal{X}|$ , then  $\sum_{x \in \mathcal{X}} \xi(x) = 0$ , and  $\|\xi\|_{TV} = \xi^+(\mathcal{X}) + \xi^-(\mathcal{X})$  denotes the total variation of  $\xi$ , where  $\xi^+ = \max\{\xi, 0\}$  and  $\xi^- = \max\{-\xi, 0\}$  stand for the positive and negative part of  $\xi$ , respectively. Therefore,

$$\sum_{x \in \mathcal{X}} \xi(x) = \sum_{x \in \mathcal{X}} \xi^+(x) - \sum_{x \in \mathcal{X}} \xi^-(x), \quad \|\xi\|_{TV} = \sum_{x \in \mathcal{X}} |\xi(x)| = \sum_{x \in \mathcal{X}} \xi^+(x) + \sum_{x \in \mathcal{X}} \xi^-(x)$$

and hence  $\sum_{x \in \mathcal{X}} \xi^+(x) \equiv \alpha/2 \equiv \sum_{x \in \mathcal{X}} \xi^-(x)$ . In addition, since

$$(2.7) \quad \sum_{x \in \mathcal{X}} \ell(x) \xi(x) = \sum_{x \in \mathcal{X}} \ell(x) \xi^+(x) - \sum_{x \in \mathcal{X}} \ell(x) \xi^-(x)$$

then (2.5) can be reformulated as follows

$$(2.8) \quad \max_{\nu \in \mathbb{B}_R(\mu)} \sum_{x \in \mathcal{X}} \ell(x) \nu(x) \longrightarrow \sum_{x \in \mathcal{X}} \ell(x) \mu(x) + \max_{\xi \in \tilde{\mathbb{B}}_R(\mu)} \sum_{x \in \mathcal{X}} \ell(x) \xi(x)$$

where  $\xi \in \tilde{\mathbb{B}}_R(\mu)$  is described by the constraints

$$(2.9) \quad \alpha \triangleq \sum_{x \in \mathcal{X}} |\xi(x)| \leq R, \quad \sum_{x \in \mathcal{X}} \xi(x) = 0, \quad 0 \leq \xi(x) + \mu(x) \leq 1, \quad \forall x \in \mathcal{X}.$$

The solution of (2.8) is obtained by first identifying the partition of  $\mathcal{X}$  into disjoint sets  $(\mathcal{X}^0, \mathcal{X} \setminus \mathcal{X}^0)$ , and then by finding upper and lower bounds on the probabilities of  $\mathcal{X}^0$  and  $\mathcal{X} \setminus \mathcal{X}^0$ , which are achievable [7].

Towards this end, define the maximum and minimum values of  $\{\ell(x) : x \in \mathcal{X}\}$  by

$$\ell_{\max} \triangleq \max_{x \in \mathcal{X}} \ell(x), \quad \ell_{\min} \triangleq \min_{x \in \mathcal{X}} \ell(x)$$

and their corresponding support sets by

$$\mathcal{X}^0 \triangleq \{x \in \mathcal{X} : \ell(x) = \ell_{\max}\}, \quad \mathcal{X}_0 \triangleq \{x \in \mathcal{X} : \ell(x) = \ell_{\min}\}.$$

For all remaining sequence,  $\{\ell(x) : x \in \mathcal{X} \setminus \mathcal{X}^0 \cup \mathcal{X}_0\}$ , and for  $1 \leq r \leq |\mathcal{X} \setminus \mathcal{X}^0 \cup \mathcal{X}_0|$ , define recursively the set of indices for which the sequence achieves its  $(k+1)^{th}$  smallest value by

$$(2.10) \quad \mathcal{X}_k \triangleq \left\{ x \in \mathcal{X} : \ell(x) = \min \left\{ \ell(\alpha) : \alpha \in \mathcal{X} \setminus \mathcal{X}^0 \cup \left( \bigcup_{j=1}^k \mathcal{X}_{j-1} \right) \right\} \right\}, \quad k \in \{1, 2, \dots, r\}$$

till all the elements of  $\mathcal{X}$  are exhausted. Further, define the corresponding values of the sequence on sets  $\mathcal{X}_k$  by

$$(2.11) \quad \ell(\mathcal{X}_k) \triangleq \min_{x \in \mathcal{X} \setminus \mathcal{X}^0 \cup \left( \bigcup_{j=1}^k \mathcal{X}_{j-1} \right)} \ell(x), \quad k \in \{1, 2, \dots, r\}$$

where  $r$  is the number of  $\mathcal{X}_k$  sets which is at most  $|\mathcal{X} \setminus \mathcal{X}^0 \cup \mathcal{X}_0|$ .

From [7] we have the following. The maximum pay-off subject to the total variation constraint is given by

$$(2.12) \quad L(\nu^*) = \ell_{\max} \nu^*(\mathcal{X}^0) + \ell_{\min} \nu^*(\mathcal{X}_0) + \sum_{k=1}^r \ell(\mathcal{X}_k) \nu^*(\mathcal{X}_k).$$

Moreover, the optimal probabilities are given by the following equations (water-filling solution).

$$(2.13a) \quad \nu^*(\mathcal{X}^0) = \mu(\mathcal{X}^0) + \frac{\alpha}{2}$$

$$(2.13b) \quad \nu^*(\mathcal{X}_0) = \left( \mu(\mathcal{X}_0) - \frac{\alpha}{2} \right)^+$$

$$(2.13c) \quad \nu^*(\mathcal{X}_k) = \left( \mu(\mathcal{X}_k) - \left( \frac{\alpha}{2} - \sum_{j=1}^k \mu(\mathcal{X}_{j-1}) \right)^+ \right)^+$$

$$(2.13d) \quad \alpha = \min \left( R, 2(1 - \mu(\mathcal{X}^0)) \right)$$

where  $R \in [0, 2]$ ,  $k \in \{1, 2, \dots, r\}$  and  $r$  is the number of  $\mathcal{X}_k$  sets which is at most  $|\mathcal{X} \setminus \mathcal{X}^0 \cup \mathcal{X}_0|$ .

The above discussion also holds for countable alphabet spaces  $(\mathcal{X}, \mathcal{U})$ . Next, we apply the above results to the minimax MCP defined by (1.12).

**3. Minimax Stochastic Control for Finite State and Control Spaces.** In this section, we investigate the infinite horizon minimax MCP defined by (1.12) for finite state and control spaces. By employing the results of Section 2, we derive dynamic programming equation (1.13) and we introduce the corresponding policy iteration algorithm.

Consider the problem of minimizing the finite horizon version of (1.11) defined by

$$(3.1) \quad J_n^*(x) = \inf_{u \in \mathcal{U}(x)} \sup_{Q(\cdot|x, u) \in \mathbf{B}_R(Q^o)(x, u)} \mathbb{E}_x^g \left\{ \sum_{k=0}^{n-1} f(x_k, u_k) \right\}.$$

Let  $V : \mathcal{X} \mapsto \mathbb{R}$  denote the value function corresponding to (3.1). Then  $V$  satisfies the dynamic programming equation [6, 19]

$$(3.2a) \quad V_n(x) = 0, \quad \forall x \in \mathcal{X}$$

$$(3.2b) \quad V_j(x) = \inf_{u \in \mathcal{U}(x)} \sup_{Q(\cdot|x, u) \in \mathbf{B}_R(Q^o)(x, u)} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} V_{j+1}(z) Q(z|x, u) \right\}, \quad j = 0, 1, \dots, n-1, \quad x \in \mathcal{X}.$$



By applying (2.1), with  $\ell(\cdot) = V_{j+1}(\cdot)$  and  $\mu(\cdot) = Q^o(\cdot|x, u)$ , then (3.2b) is equivalent to the dynamic programming equation

$$(3.3) \quad V_j(x) = \inf_{u \in \mathcal{U}(x)} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} V_{j+1}(z) Q^o(z|x, u) + \frac{R}{2} \left( \sup_{z \in \mathcal{X}} V_{j+1}(z) - \inf_{z \in \mathcal{X}} V_{j+1}(z) \right) \right\}.$$

Moreover, by applying (2.13) with  $\nu^*(\cdot) = Q^*(\cdot|x, u)$ , where  $Q^*(\cdot|x, u)$ ,  $(x, u) \in \mathbb{K}$  is the maximizing conditional distribution and  $\mu(\cdot) = Q^o(\cdot|x, u)$ ,  $(x, u) \in \mathbb{K}$ , then (3.2b) is equivalent to

$$(3.4) \quad V_j(x) = \inf_{u \in \mathcal{U}(x)} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} V_{j+1}(z) Q^*(z|x, u) \right\}.$$

Define  $\bar{V}_j(x) = V_{n-j}(x)$ . Then from (3.2b),  $\bar{V}_j(\cdot)$  satisfies the equation

$$(3.5) \quad \begin{aligned} \bar{V}_j(x) = \inf_{u \in \mathcal{U}(x)} \sup_{Q(\cdot|x, u) \in \mathbf{B}_R(Q^o)(x, u)} & \\ \left\{ f(x, u) + \sum_{z \in \mathcal{X}} \bar{V}_{j-1}(z) Q(z|x, u) \right\}, & \quad j = 0, 1, \dots, n-1. \end{aligned}$$

We rewrite (3.5) as follows.

$$(3.6) \quad \begin{aligned} & \bar{V}_j(x) + \frac{1}{j} \bar{V}_j(x) \\ &= \inf_{u \in \mathcal{U}(x)} \sup_{Q(\cdot|x, u) \in \mathbf{B}_R(Q^o)(x, u)} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} Q(z|x, u) \left( \bar{V}_{j-1}(z) + \frac{1}{j} \bar{V}_j(x) \right) \right\}. \end{aligned}$$

Next, we introduce the following standard assumption [14].

ASSUMPTION 3.1. *There exists a pair  $(V(\cdot), J^*)$ ,  $V : \mathcal{X} \mapsto \mathbb{R}$  and  $J^* \in \mathbb{R}$ , such that*

$$(3.7) \quad \lim_{j \rightarrow \infty} (\bar{V}_j(x) - jJ^*) = V(x), \quad \forall x \in \mathcal{X}.$$

Under Assumption 3.1, then

$$(3.8) \quad \lim_{j \rightarrow \infty} \frac{1}{j} \bar{V}_j(x) = J^*, \quad \forall x \in \mathcal{X}$$

and the limit does not depend on  $x \in \mathcal{X}$ . In addition, by taking the supremum with respect to  $x \in \mathcal{X}$  on both sides of (3.7), by virtue of the finite cardinality of  $\mathcal{X}$ , we can exchange the limit and the supremum to obtain

$$(3.9) \quad \lim_{j \rightarrow \infty} \sup_{x \in \mathcal{X}} (\bar{V}_j(x) - jJ^*) = \sup_{x \in \mathcal{X}} \lim_{j \rightarrow \infty} (\bar{V}_j(x) - jJ^*) = \sup_{x \in \mathcal{X}} V(x).$$

By Assumption 3.1 and by (3.8) we have the following identities.

$$\begin{aligned}
& J^* + V(x) \\
&= \lim_{j \rightarrow \infty} \left( \frac{1}{j} \bar{V}_j(x) + (\bar{V}_j(x) - jJ^*) \right) \\
&\stackrel{(a)}{=} \lim_{j \rightarrow \infty} \inf_{u \in \mathcal{U}(x)} \sup_{Q(\cdot|x, u) \in \mathbf{B}_R(Q^o)(x, u)} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} Q(z|x, u) \left( \bar{V}_{j-1}(z) + \frac{1}{j} \bar{V}_j(x) \right) - jJ^* \right\} \\
&\stackrel{(b)}{=} \lim_{j \rightarrow \infty} \inf_{u \in \mathcal{U}(x)} \left\{ f(x, u) - jJ^* + \sum_{z \in \mathcal{X}} Q^o(z|x, u) \left( \bar{V}_{j-1}(z) + \frac{1}{j} \bar{V}_j(x) \right) \right. \\
&\quad \left. + \frac{R}{2} \left( \sup_{z \in \mathcal{X}} \left( \bar{V}_{j-1}(z) + \frac{1}{j} \bar{V}_j(x) \right) - \inf_{z \in \mathcal{X}} \left( \bar{V}_{j-1}(z) + \frac{1}{j} \bar{V}_j(x) \right) \right) \right\} \\
&\stackrel{(c)}{=} \lim_{j \rightarrow \infty} \inf_{u \in \mathcal{U}(x)} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} Q^o(z|x, u) (\bar{V}_{j-1}(z) - (j-1)J^* + \frac{1}{j} \bar{V}_j(x) - J^*) \right. \\
&\quad \left. + \frac{R}{2} \left( \sup_{z \in \mathcal{X}} (\bar{V}_{j-1}(z) - jJ^*) - \inf_{z \in \mathcal{X}} (\bar{V}_{j-1}(z) - jJ^*) \right) \right\}
\end{aligned}$$

where

(a) is obtained by using (3.6);

(b) is obtained by using the equivalent formulation (3.3);

(c) is obtained by adding and subtracting  $J^*(1 + j\frac{R}{2})$ .

Since  $\mathcal{U}$  and  $\mathcal{X}$  are of finite cardinality we can interchange the limit and the minimization and maximization operations, to arrive to the following dynamic programming equation.

$$(3.10) \quad J^* + V(x) = \min_{u \in \mathcal{U}(x)} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} Q^o(z|x, u) V(z) + \frac{R}{2} \left( \sup_{z \in \mathcal{X}} V(z) - \inf_{z \in \mathcal{X}} V(z) \right) \right\}.$$

Clearly, by (2.1), dynamic programming equation (3.10) is equivalently expressed as follows.

$$(3.11) \quad J^* + V(x) = \min_{u \in \mathcal{U}(x)} \max_{Q(\cdot|x, u) \in \mathbf{B}_R(Q^o)(x, u)} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} Q(z|x, u) V(z) \right\}.$$

Next, we state the first main Theorem of this section.

**THEOREM 3.2.** *Suppose  $\mathcal{X}$  and  $\mathcal{U}$  are of finite cardinality and Assumption 3.1 holds. If there exists a solution  $(V, J^*)$  to the dynamic programming equation (3.10), and  $g^*$  is a stationary policy such that  $g^*(x)$  attains the minimum in the right-hand side of (3.10) for every  $x$ , then  $g^*$  is an optimal policy and  $J^*$  is the minimum average cost.*

*Proof.* Let  $g \in G$  be any policy and  $u \in \mathcal{U}(x)$ . Since  $(V, J^*)$  satisfies the dynamic programming equation (3.10), which is equivalent to (3.11), and by the definition of  $g^*$  then

$$\begin{aligned}
(3.12) \quad & f(x, u) + \sum_{z \in \mathcal{X}} Q^o(z|x, u) V(z) + \frac{R}{2} \left( \max_{z \in \mathcal{X}} V(z) - \min_{z \in \mathcal{X}} V(z) \right) \\
&= \max_{Q(\cdot|x, u) \in \mathbf{B}_R(Q^o)(x, u)} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} Q(z|x, u) V(z) \right\} \\
&\geq \max_{Q(\cdot|x, g^*(x)) \in \mathbf{B}_R(Q^o)(x, g^*(x))} \left\{ f(x, g^*(x)) + \sum_{z \in \mathcal{X}} Q(z|x, g^*(x)) V(z) \right\} \\
&= J^* + V(x).
\end{aligned}$$

Denoting the maximization with respect to  $Q(\cdot|x, u)$  in (3.12) by  $Q^*(\cdot|x, u)$  and the corresponding expectation by  $\mathbb{E}^{g, Q^*}$ , and taking expectation on both sides of (3.12), we have

$$\begin{aligned} \mathbb{E}^{g, Q^*}(f(x_j, u_j)) &\geq J^* + \mathbb{E}^{g, Q^*}(V(x_j)) - \mathbb{E}^{g, Q^*}\left(\sum_{z \in \mathcal{X}} Q^*(z|x_j, u_j)V(z)\right) \\ (3.13) \quad &= J^* + \mathbb{E}^{g, Q^*}(V(x_j)) - \mathbb{E}^{g, Q^*}(V(x_{j+1})). \end{aligned}$$

Then, from (1.11) we have that for all  $g \in G$ ,

$$\begin{aligned} J(\pi) &\geq \liminf_{j \rightarrow \infty} \left( \frac{1}{j} \sum_{k=0}^{j-1} \mathbb{E}^{g, Q^*}(f(x_k, u_k)) \right) \\ &\stackrel{(a)}{\geq} \liminf_{j \rightarrow \infty} \left( J^* + \frac{1}{j} \left( \mathbb{E}^{g, Q^*}(V(x_0)) - \mathbb{E}^{g, Q^*}(V(x_j)) \right) \right) \\ &\stackrel{(b)}{=} J^* \end{aligned}$$

where

(a) is obtained by using (3.13);

(b) is obtained because the last term vanishes as  $j \rightarrow \infty$ .

Thus,  $J^* \leq \inf_{g \in G} J(g, x)$ . However, when  $g$  is replaced by  $g^*$  equality holds throughout, and as a result  $g^*$  is optimal, that is,  $J^* = J^*(x) = \inf_{g \in G} J(g, x)$ ,  $g^* \in G$  is an average cost optimal policy and  $J^*$  is the value.  $\square$

**3.1. Existence.** Dynamic programming equation (3.10) and hence Theorem 3.2, are valid under Assumption 3.1. Here, we characterize the solution of the infinite horizon minimax average cost MCM, under the standard irreducibility condition, on the nominal transition probabilities of the controlled process. First, we introduce some notation.

Identify the state space  $\mathcal{X}$  by  $\mathcal{X} = \{x_1, x_2, \dots, x_{|\mathcal{X}|}\}$  consisting of  $|\mathcal{X}|$  elements. Then, any function  $V : \mathcal{X} \rightarrow \mathbb{R}$  may be represented by a vector in  $\mathbb{R}^{|\mathcal{X}|}$ , as follows.

$$V = (V(x_1) \quad \dots \quad V(x_{|\mathcal{X}|}))^T \in \mathbb{R}^{|\mathcal{X}|}.$$

Any stationary control policy  $g \in G_{SM}$ ,  $g : \mathcal{X} \rightarrow \mathbb{R}$ , may also be identified with a  $g \in \mathbb{R}^{|\mathcal{X}|}$ . For any  $g$ , let  $Q(g) \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{X}|}$  defined by  $Q(g)_{ij} = P(x_{t+1} = x_i | x_t = x_j, u_t = g(x_j))$  and

$$f(g) = (f(x_1, g(x_1)) \quad \dots \quad f(x_{|\mathcal{X}|}, g(x_{|\mathcal{X}|})))^T \in \mathbb{R}^{|\mathcal{X}|}.$$

Let  $q_0 \in \mathbb{R}^{|\mathcal{X}|}$  be defined by  $q_0(x_i) \triangleq P(\{x_0 = x_i\})$ ,  $i = 1, \dots, |\mathcal{X}|$  and  $e \triangleq (1, \dots, 1)^T \in \mathbb{R}^{|\mathcal{X}|}$ .

The maximization of the expected  $n$ -stage cost, for a fixed  $q_0(x) \in \mathbb{R}^{|\mathcal{X}|}$ , is given by<sup>4</sup>

$$\begin{aligned} J_n(g, q_0) &\triangleq J_n(g, x) q_0^T(x) = \max_{Q(\cdot|x, u) \in \mathbf{B}_R(Q^o)(x, u)} \mathbb{E}^g \left\{ \sum_{k=0}^{n-1} f(x_k, u_k) \right\} \\ (3.14) \quad &= \max_{Q(\cdot|x, u) \in \mathbf{B}_R(Q^o)(x, u)} \left\{ \sum_{k=0}^{n-1} q_0^T Q(g)^k f(g) \right\} \\ &= \max_{Q(\cdot|x, u) \in \mathbf{B}_R(Q^o)(x, u)} q_0^T \left\{ \sum_{k=0}^{n-1} Q(g)^k \right\} f(g). \end{aligned}$$

<sup>4</sup>The notation  $J_n(g, q_0)$  means that  $q_0(x)$  is fixed instead of  $x_0 = x$ .

With  $Q^*(\cdot|x, u)$  denoting the maximizing conditional distribution, then (3.14) is equivalent

$$\max_{Q(\cdot|x, u) \in \mathbf{B}_R(Q^o)(x, u)} q_0^T \left\{ \sum_{k=0}^{n-1} Q(g)^k \right\} f(g) = q_0^T \left\{ \sum_{k=0}^{n-1} Q^*(g)^k \right\} f(g).$$

Hence, the maximizing average cost per unit-time is given by

$$\begin{aligned} J(g, q_0) &= \limsup_{n \rightarrow \infty} \max_{Q(\cdot|x, u) \in \mathbf{B}_R(Q^o)(x, u)} \frac{1}{n} \mathbb{E}^g \left\{ \sum_{k=0}^{n-1} f(x_k, u_k) \right\} \\ &= \limsup_{n \rightarrow \infty} \frac{1}{n} q_0^T \left\{ \sum_{k=0}^{n-1} Q^*(g)^k \right\} f(g). \end{aligned}$$

Since  $q_0 \in \mathbb{R}^{|\mathcal{X}|}$  and  $f(g) \in \mathbb{R}^{|\mathcal{X}|}$  are independent of  $n$ , we only need to investigate the conditions under which the following limit exists

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} Q^*(g)^k.$$

The next Lemma follows directly from [14, Lemma 5.4].

LEMMA 3.3. *If  $Q^* \in \mathbb{R}_+^{|\mathcal{X}| \times |\mathcal{X}|}$  is a stochastic matrix, then the Cesaro limit*

$$(3.15) \quad \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} (Q^*)^k = Q_1^*$$

*always exist. The matrix  $Q_1^* \in \mathbb{R}_+^{|\mathcal{X}| \times |\mathcal{X}|}$  is a stochastic matrix and it is the solution of the equation*

$$(3.16) \quad Q_1^* Q^* = Q_1^*.$$

In view of Lemma 3.3, the maximization of the average cost per unit-time of a stationary Markov control policy is given by

$$(3.17) \quad J(g, q_0) = q_0^T Q_1^*(g) f(g)$$

where  $Q_1^*(g)$  and  $Q^*(g)$  are related by (3.16). We recall the following definition of reducible stochastic matrix from [14, page 44].

DEFINITION 3.4. *A stochastic matrix  $P \in \mathbb{R}_+^{|\mathcal{X}| \times |\mathcal{X}|}$  is said to be reducible if by row and column permutations it can be placed into block upper-triangular form*

$$P = \begin{pmatrix} P_1 & P_2 \\ 0 & P_3 \end{pmatrix}, \quad \text{where } P_1, P_2 \text{ are square matrices.}$$

*A stochastic matrix which is not reducible is said to be irreducible.*

Next, we recall the following Lemma from [14, Lemma 5.7].

LEMMA 3.5. *Let  $Q^* \in \mathbb{R}_+^{|\mathcal{X}| \times |\mathcal{X}|}$  be an irreducible stochastic matrix. Then, there exists a unique vector  $q$  such that*

$$Q^* q = q, \quad e^T q = 1, \quad q(x_i) > 0 \text{ for all } x_i \in \mathcal{X}.$$

Moreover, the matrix  $Q_1^*$  associated with  $Q^*$  in (3.16) has all rows equal to  $q$ .

Note that, (3.17) depends on the probability distribution  $q_0$  of the initial state. However, if  $Q_1^*$  is assumed to be an irreducible stochastic matrix, by Lemma 3.5

$$(3.18) \quad J(g, q_0) = q_0^T Q_1^*(g) f(g) = q(g)^T f(g) \equiv J(g)$$

where  $q(g)$  is the unique invariant probability distribution, that is,  $Q^*(g)q(g) = q(g)$ , and the average cost per unit-time  $J(g, q_0) \equiv J(g)$  is independent of the initial distribution. Hence, for the remainder of this section, we will assume that for every stationary Markov control policy  $g \in G_{SM}$ , the stochastic matrix  $Q^*(g)$  is irreducible. The next proposition summarizes the above results.

**PROPOSITION 3.6.** [20] *Let  $g \in G_{SM}$  be a stationary Markov control policy,  $g : \mathcal{X} \mapsto \mathcal{U}$  and assume that  $Q^*(g) \in \mathbb{R}_+^{|\mathcal{X}| \times |\mathcal{X}|}$  is irreducible. Then the following hold.*

(a) *There exists a unique  $q(g) \in \mathbb{R}_+^{|\mathcal{X}|}$  such that*

$$(3.19) \quad Q^*(g)q(g) = q(g), \quad e^T q = 1.$$

(b) *The average cost per unit-time associated with the control policy  $g \in G_{SM}$  is*

$$(3.20) \quad J(g) = q(g)^T f(g).$$

(c) *There exists a  $V(g) \in \mathbb{R}^{|\mathcal{X}|}$  such that*

$$(3.21) \quad J(g)e + V(g) = f(g) + Q^*(g)V(g).$$

*Proof.* Part (a) and (b) follows from Lemma (3.5) and the discussion above it. For part (c) see [20].  $\square$

**LEMMA 3.7.** *Assume the following hold.*

1. *For any stationary control policy  $g \in G_{SM}$ ,  $Q^*(g) \in \mathbb{R}_+^{|\mathcal{X}| \times |\mathcal{X}|}$  is irreducible.*
2. *There exists a  $g^* \in G_{SM}$  such that*

$$J^* = \inf_{g \in G_{SM}} J(g).$$

*Then there exists an  $(V(g^*, \cdot), J^*)$ ,  $V(g^*, \cdot) : \mathcal{X} \mapsto \mathbb{R}$  and  $J^* \in \mathbb{R}$  which is a solution to the dynamic programming equation*

$$J^* + V(g^*, x) = \min_{u \in \mathcal{U}} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} Q^*(z|x, u) V(g^*, z) \right\}.$$

*Proof.* By Proposition 3.6 (c), there exists a  $V(g^*, \cdot) : \mathcal{X} \mapsto \mathbb{R}$  and  $J^*$  such that for all  $x \in \mathcal{X}$

$$(3.22) \quad J^* + V(g^*, x) = f(x, g^*(x)) + \sum_{z \in \mathcal{X}} Q^*(z|x, g^*(x)) V(g^*, z).$$

Then, for all  $x \in \mathcal{X}$

$$J^* + V(g^*, x) \geq \min_{u \in \mathcal{U}} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} Q^*(z|x, u) V(g^*, z) \right\}.$$

Define  $g_1 : \mathcal{X} \mapsto \mathcal{U}$  as

$$g_1(x) = \operatorname{argmin}_{u \in \mathcal{U}} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} Q^*(z|x, u) V(g^*, z) \right\}.$$

Suppose that for some  $x_2 \in \mathcal{X}$  strict inequality holds in (3.22), then

$$(3.23) \quad J^* + V(g^*, x) > \min_{u \in \mathcal{U}} \left\{ f(x_2, u) + \sum_{z \in \mathcal{X}} Q^*(z|x_2, u) V(g^*, z) \right\}.$$

Multiplying (3.23) by  $q(g_1)(x_0) > 0$  and summing over  $x_0 \in \mathcal{X}$  yields

$$\begin{aligned} & J^* + \sum_{x_0 \in \mathcal{X}} q(g_1)(x_0) V(g^*, x_0) \\ & > \min_{u \in \mathcal{U}} \left\{ \sum_{x_0 \in \mathcal{X}} q(g_1)(x_0) f(x_0, u) + \sum_{x_0 \in \mathcal{X}} q(g_1)(x_0) \sum_{z \in \mathcal{X}} Q^*(z|x_0, u) V(g^*, z) \right\} \\ & = \sum_{x_0 \in \mathcal{X}} q(g_1)(x_0) f(x_0, g_1(x_0)) + \sum_{x_0 \in \mathcal{X}} q(g_1)(x_0) \sum_{z \in \mathcal{X}} Q^*(z|x_0, g_1(x_0)) V(g^*, z) \\ & = J(g_1) + \sum_{z \in \mathcal{X}} q(g_1) V(g^*, z), \quad \text{by Proposition 3.6 (a)} \end{aligned}$$

which gives  $J^* > J(g_1)$ , contradicting assumption 2. Hence, equality holds in (3.22), for every  $x \in \mathcal{X}$ .  $\square$

Next, we state the second main Theorem of this section.

**THEOREM 3.8.** *Assume that for all stationary Markov control policies  $g \in G_{SM}$ , and for a given total variation parameter  $R \in [0, R_{\max}] \subset [0, 2]$ , the maximizing transition matrix  $Q^*(g)$  is irreducible. Then the following hold.*

(a) *There exists a solution  $(V, J^*)$ ,  $V : \mathcal{X} \mapsto \mathbb{R}$ ,  $J^* \in \mathbb{R}$  to the dynamic programming equation*

$$(3.24) \quad J^* + V(x) = \min_{u \in \mathcal{U}} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} Q^*(z|x, u) V(z) \right\}$$

*or, to the equivalent dynamic programming equation*

$$(3.25) \quad J^* + V(x) = \min_{u \in \mathcal{U}} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} Q^o(z|x, u) V(z) + \frac{R}{2} \left( \max_{z \in \mathcal{X}} V(z) - \min_{z \in \mathcal{X}} V(z) \right) \right\}$$

*where  $\max_{z \in \mathcal{X}} V(z)$  denotes component-wise maximum and similarly for the minimum. The maximizing conditional distribution  $Q^*(\cdot|x, u)$ ,  $(x, u) \in \mathbb{K}$  is given by (2.13), where  $\nu^*(\cdot)$ ,  $\mu(\cdot)$  and  $\ell(\cdot)$  are replaced by  $Q^*(\cdot|x, u)$ ,  $Q^o(\cdot|x, u)$  and  $V(\cdot)$ , respectively, i.e.,*

$$(3.26a) \quad Q^*(\mathcal{X}^+|x, u) = Q^o(\mathcal{X}^+|x, u) + \frac{\alpha}{2}$$

$$(3.26b) \quad Q^*(\mathcal{X}^-|x, u) = \left( Q^o(\mathcal{X}^-|x, u) - \frac{\alpha}{2} \right)^+$$

$$(3.26c) \quad Q^*(\mathcal{X}_k|x, u) = \left( Q^o(\mathcal{X}_k|x, u) - \left( \frac{\alpha}{2} - \sum_{j=1}^k Q^o(\mathcal{X}_{k-1}|x, u) \right)^+ \right)^+$$

$$(3.26d) \quad \alpha = \min \left( R, 2(1 - Q^o(\mathcal{X}^+|x, u)) \right)$$

and

$$(3.27a) \quad \mathcal{X}^+ \triangleq \left\{ x \in \mathcal{X} : V(x) = \max\{V(x) : x \in \mathcal{X}\} \right\}$$

$$(3.27b) \quad \mathcal{X}^- \triangleq \left\{ x \in \mathcal{X} : V(x) = \min\{V(x) : x \in \mathcal{X}\} \right\}$$

$$(3.27c) \quad \mathcal{X}_k \triangleq \left\{ x \in \mathcal{X} : V(x) = \min \left\{ V(\alpha) : \alpha \in \mathcal{X} \setminus \mathcal{X}^0 \cup \left( \bigcup_{j=1}^k \mathcal{X}_{j-1} \right) \right\} \right\}$$

where  $k = 1, 2, \dots, r$  (see Section 2.1).

(b) If  $g^*(x)$  attains the minimum in (3.24) or equivalently in (3.25) for every  $x$ , then  $g^*$  is an average cost optimal policy.

(c) The minimum average cost is  $J^*$ .

*Proof.* Theorem 3.8 is obtained by combining Theorem 3.2 and Lemma 3.7 and by applying the results of Section 2.  $\square$

The main observation is that in specific applications one may employ either dynamic programming equation (3.24) or (3.25).

**3.1.1. Policy Iteration Algorithm.** In this section, we provide a modified version of the classical policy iteration algorithm for average cost dynamic programming [14, 20]. From part (a) of Theorem 3.8, the policy evaluation and policy improvement steps of a policy iteration algorithm must be performed using the maximizing conditional distribution obtained under total variation distance ambiguity constraint. Moreover, one needs to guarantee that for the given total variation parameter  $R$ , the corresponding maximizing matrix  $Q^*$  is irreducible, otherwise, Algorithm 3.9 may not be sufficient to give the optimal policy and the minimum cost. In general,  $R \in [0, R_{\max}] \subseteq [0, 2]$ , and  $R_{\max}$  is strictly less than 2. This generality will be discussed in Section 4 for general Borel spaces.

ALGORITHM 3.9. (Policy iteration)

1. Let  $m = 0$  and select an arbitrary stationary Markov control policy  $g_0 : \mathcal{X} \mapsto \mathcal{U}$ .
2. (Policy Evaluation) Solve the equation

$$(3.28) \quad J_{Q^o}(g_m)e + V_{Q^o}(g_m) = f(g_m) + Q^o(g_m)V_{Q^o}(g_m)$$

for  $J_{Q^o}(g_m) \in \mathbb{R}$  and  $V_{Q^o}(g_m) \in \mathbb{R}^{|\mathcal{X}|}$ . Identify the support sets of (3.28) using (3.27), and construct the matrix  $Q^*(g_m)$  using (3.26). Solve the equation

$$(3.29) \quad J_{Q^*}(g_m)e + V_{Q^*}(g_m) = f(g_m) + Q^*(g_m)V_{Q^*}(g_m)$$

for  $J_{Q^*}(g_m) \in \mathbb{R}$  and  $V_{Q^*}(g_m) \in \mathbb{R}^{|\mathcal{X}|}$ .

3. (Policy Improvement) Let

$$(3.30) \quad g_{m+1} = \arg \min_{g \in \mathbb{R}^{|\mathcal{X}|}} \left\{ f(g) + Q^*(g)V_{Q^*}(g_m) \right\}.$$

4. If  $g_{m+1} = g_m$ , let  $g^* = g_m$ ; else let  $m = m + 1$  and return to step 2.

In Section 5.1, we illustrate how policy iteration algorithm for infinite horizon average cost dynamic programming is implemented through an example.

**3.1.2. Limitations.** Part (a) of Theorem 3.8, indicates that for a stationary Markov control policy  $g \in G_{SM}$ , and for an irreducible stochastic matrix  $Q^*$  there exists a solution to the dynamic programming equation (3.24). Moreover, the maximizing stochastic matrix  $Q^*$

which is given by (3.26), is calculated based on the support sets (3.27), the nominal stochastic matrix  $Q^o$ , and the value of the total variation parameter  $R \in [0, R_{\max}]$ . Hence, in order to apply policy iteration algorithm for average-cost dynamic programming one needs to know in advance that, for a given total variation parameter  $R \in [0, 2]$ , and an irreducible nominal stochastic matrix  $Q^o$ , the maximizing stochastic matrix  $Q^*$  is also irreducible. Otherwise, policy iteration algorithm may not be sufficient to give the optimal policy and the minimum cost. In particular, as we show next, if the irreducibility condition is not satisfied then the policy iteration algorithm need not have a unique solution.

As an example (inspired by [16]), consider the stochastic control system shown in Fig. 3.1, with state-space  $\mathcal{X} = \{1, 2, 3\}$  and control set  $\mathcal{U} = \{u_1, u_2\}$ . Let the nominal transition prob-

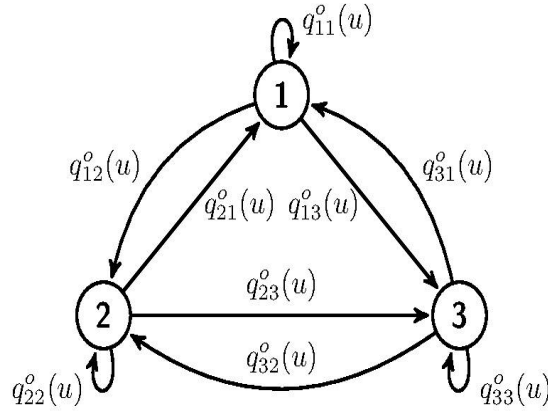


FIG. 3.1. Nominal Stochastic Control System.

ability under controls  $u_1$  and  $u_2$  to be given by

$$(3.31) \quad Q^o(u_1) = \frac{1}{9} \begin{pmatrix} 0 & 5 & 4 \\ 0 & 9 & 0 \\ 0 & 0 & 9 \end{pmatrix}, \quad Q^o(u_2) = \frac{1}{9} \begin{pmatrix} 2 & 7 & 0 \\ 3 & 6 & 0 \\ 8 & 0 & 1 \end{pmatrix}.$$

The cost function under each state and action is given by

$$f(1, u_1) = 2, f(2, u_1) = 1, f(3, u_1) = 3, f(1, u_2) = 0.5, f(2, u_2) = 3, f(3, u_2) = 0.$$

Clearly, from (3.31), this control system the nominal transition probability matrix, under both controls, is reducible, since the system under controls  $u_1$  and  $u_2$  contains more than one communication class<sup>5</sup>. Using policy iteration Algorithm 3.9 with initial policies  $g_0(1) = g_0(2) = g_0(3) = u_1$ , the optimality equation (3.28) for this system may be written as

$$J_{Q^o} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + V_{Q^o}(g_0) = \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix} + \frac{1}{9} \begin{pmatrix} 0 & 5 & 4 \\ 0 & 9 & 0 \\ 0 & 0 & 9 \end{pmatrix} V_{Q^o}(g_0)$$

and hence

$$\begin{aligned} J_{Q^o} + V_{Q^o}(g_0, 1) &= 2 + \frac{5}{9}V_{Q^o}(g_0, 2) + \frac{4}{9}V_{Q^o}(g_0, 3) \\ J_{Q^o} + V_{Q^o}(g_0, 2) &= 1 + V_{Q^o}(g_0, 2) \implies J_{Q^o} = 1 \\ J_{Q^o} + V_{Q^o}(g_0, 3) &= 3 + V_{Q^o}(g_0, 3) \implies J_{Q^o} = 3. \end{aligned}$$

<sup>5</sup>States  $i$  and  $j$  belong to the same communication class if and only if each of these states can reach and be reached by the other.



The second and third equations show that the system is inconsistent, and hence, the policy iteration algorithm fails to give the optimal policy and the minimum cost.

Moreover, even if  $Q^o$  is an irreducible stochastic matrix, as the value of total variation parameter  $R$  increases the maximizing stochastic matrix  $Q^*(R)$ , eventually, will be transformed into a reducible stochastic matrix. Hence, our proposed method for solving minimax stochastic control problem with average cost is valid only for a specific range of values of total variation parameter, in  $R \in [0, R_{\max}] \subseteq [0, 2]$ . In particular, if  $Q^o$  is an irreducible stochastic matrix then, for any given partition of the state-space, there exists an  $R_{\max} \in [0, 2)$  for which we distinguish the following two cases:

- (a) for  $0 \leq R < R_{\max}$ ,  $Q^*$  is an irreducible stochastic matrix. Theorem 3.8 is valid and policy iteration algorithm gives the optimal policy and the minimum cost.
- (b) for  $R \geq R_{\max}$ ,  $Q^*$  is a reducible stochastic matrix. Theorem 3.8 is not valid and policy iteration algorithm need not have a solution.

**REMARK 3.10.** Consider  $R \geq R_{\max}$ . Then, an extended solution through a reduced dimensional state-space may be obtained as follows. Due to the water-filling behavior of maximizing conditional distribution (3.26), columns of  $Q^*$  which correspond to states belonging to  $\mathcal{X} \setminus \mathcal{X}^0$ , become columns with all zero's, as total variation parameter  $R$  increases. Whenever an all zero column appears, one can remove the corresponding state of that column, and hence  $Q^*$  will be transformed back into an irreducible stochastic matrix of reduced order.

**4. Minimax Stochastic Control for Borel Spaces.** In this section, we derive the general dynamic programming equation for Borel spaces  $(\mathcal{X}, \mathcal{U})$  which solves the MDP for all values of  $R \in [0, 2]$ . In addition, we derive a generalized policy iteration algorithm corresponding to the generalized dynamic programming equations when the state and control spaces are of finite dimension. Note that, throughout this section we again suppose that Assumption 1.4 holds.

**4.1. General Dynamic Programming.** Throughout this section it is assumed that Assumptions 1.4 hold. The characterization of optimal policies for the minimax MCP defined by (1.12), will be based on the concept of a canonical triplet adopted to the current formulation (see [12]).

Consider the MCM (1.1), where  $(\mathcal{X}, \mathcal{U})$  are Borel spaces, and let  $h : \mathcal{X} \mapsto \mathbb{R}$  be a bounded, continuous and non-negative function. Denote the expected  $n$ -stage cost, with a terminal cost  $h$ , policy  $g$ , and  $x_0 = x$ , by  $J_0(g, Q, x, h) = h(x)$ , and for  $n \geq 1$ , by

$$(4.1) \quad J_n(g, Q, x, h) = \mathbb{E}_x^g \left\{ \sum_{k=0}^{n-1} f(x_k, u_k) + h(x_n) \right\} = J_n(g, Q, x) + \mathbb{E}_x^g \{ h(x_n) \}$$

with  $J_n(g, Q, x) = J_n(g, Q, x, 0)$ . The corresponding maximizing expected  $n$ -stage cost is given by

$$(4.2) \quad J_n(g, x, h) = \sup_{Q(\cdot|x, u) \in \mathbf{B}_R(Q^o)(x, u)} \mathbb{E}_x^g \left\{ \sum_{k=0}^{n-1} f(x_k, u_k) + h(x_n) \right\} \\ = \mathbb{E}_x^{g, Q^*} \left\{ \sum_{k=0}^{n-1} f(x_k, u_k) + h(x_n) \right\} = J_n(g, x) + \mathbb{E}_x^{g, Q^*} \{ h(x_n) \}$$

with  $J_n(g, x) = J_n(g, x, 0)$ , where  $Q^*(\cdot|x, u)$  is the maximizing distribution. Then,

$$(4.3) \quad J_n^*(x, h) = \inf_{g \in G} J_n(g, x, h); \quad J_n^*(x) = \inf_{g \in G} J_n(g, x, h), \quad \text{if } h(\cdot) = 0.$$

Throughout this section it is assumed that there exists a policy  $g \in G$  and an initial state  $x \in \mathcal{X}$  such that  $J(g, x) < \infty$  (i.e., see (1.11)). The definition of a canonical triplet is introduced next, following [12, 21] with a slight variation, to account for the extra terms, which enter the dynamic programming equation.

**DEFINITION 4.1.** *Let  $\rho$  and  $h$  be real-valued, bounded, continuous, non-negative, measurable functions on  $\mathcal{X}$  and  $\varphi \in \mathbb{F}$  a given selector. Then  $(\rho, h, \varphi)$  is said to be a canonical triplet if*

$$(4.4) \quad J_n(g^\infty, x, h) = J_n^*(x, h) = n\rho(x) + h(x), \quad \forall x \in \mathcal{X}, n = 0, 1, \dots$$

A selector  $\varphi \in \mathbb{F}$  (of a stationary policy  $g^\infty \in G_{SM}$ ) is called canonical if it is an element of some canonical triplet.

Note that with the appropriate choice of  $h$  as the terminal cost the policy  $g^\infty$  is optimal for the  $n$ -stage problem for all  $n = 0, 1, \dots$ . The following Theorem characterizes the canonical triplets for the minimax problem, with respect to the new dynamic programming equation.

**THEOREM 4.2.** *Suppose the supremum and infimum of  $h(\cdot)$  and  $\rho(\cdot)$  over  $\mathcal{X}$  is non-empty. Then  $(\rho, h, \varphi)$  is a canonical triplet if and only if, for every  $x \in \mathcal{X}$ , the following hold.*

- (a)  $\rho(x) = \inf_{u \in \mathcal{U}(x)} \left\{ \int_{\mathcal{X}} \rho(z) Q^o(dz|x, u) + \frac{R}{2} \left( \sup_{z \in \mathcal{X}} \rho(z) - \inf_{z \in \mathcal{X}} \rho(z) \right) \right\}$
- (b)  $\rho(x) + h(x) = \inf_{u \in \mathcal{U}(x)} \left\{ f(x, u) + \int_{\mathcal{X}} h(z) Q^o(dz|x, u) + \frac{R}{2} \left( \sup_{z \in \mathcal{X}} h(z) - \inf_{z \in \mathcal{X}} h(z) \right) \right\}$
- (c)  $\varphi(x) \in \mathcal{U}(x)$  attains the minimum in both (a) and (b), that is,

$$(4.5) \quad \rho(x) = \int_{\mathcal{X}} \rho(z) Q^o(dz|x, \varphi) + \frac{R}{2} \left( \sup_{z \in \mathcal{X}} \rho(z) - \inf_{z \in \mathcal{X}} \rho(z) \right)$$

$$(4.6) \quad \rho(x) + h(x) = f(x, \varphi) + \int_{\mathcal{X}} h(z) Q^o(dz|x, \varphi) + \frac{R}{2} \left( \sup_{z \in \mathcal{X}} h(z) - \inf_{z \in \mathcal{X}} h(z) \right)$$

or, equivalently,  $(\rho, h, \varphi)$  is a canonical triplet if and only if for every  $x \in \mathcal{X}$  the following hold.

- (a')  $\rho(x) = \inf_{u \in \mathcal{U}(x)} \sup_{Q(\cdot|x, u) \in \mathbf{B}_R(Q^o)(x, u)} \int_{\mathcal{X}} \rho(z) Q(dz|x, u)$
  - (b')  $\rho(x) + h(x) = \inf_{u \in \mathcal{U}(x)} \sup_{Q(\cdot|x, u) \in \mathbf{B}_R(Q^o)(x, u)} \left\{ f(x, u) + \int_{\mathcal{X}} h(z) Q(dz|x, u) \right\}$
  - (c')  $\varphi(x) \in \mathcal{U}(x)$  attains the minimum in (a') and (b'), that is,
- $$(4.7) \quad \rho(x) = \sup_{Q(\cdot|x, u) \in \mathbf{B}_R(Q^o)(x, u)} \int_{\mathcal{X}} \rho(z) Q(dz|x, \varphi)$$
- $$(4.8) \quad \rho(x) + h(x) = \sup_{Q(\cdot|x, u) \in \mathbf{B}_R(Q^o)(x, u)} \left\{ f(x, \varphi) + \int_{\mathcal{X}} h(z) Q(dz|x, \varphi) \right\}$$

Note that, if  $(\rho, h, \varphi)$  is a canonical triplet, then so is  $(\rho, h + N, \varphi)$  for any constant  $N$ . Next we proceed with the proof of Theorem 4.2.

*Proof.* (Necessity). Suppose that  $(\rho, h, \varphi)$  is a canonical triplet, i.e., (4.4) holds  $\forall x \in \mathcal{X}$  and  $n \geq 0$ . From the analog of dynamic programming equation (3.3) of Borel spaces, we have that

$$(4.9) \quad V_j(x) = \inf_{u \in \mathcal{U}(x)} \left\{ f(x, u) + \int_{\mathcal{X}} V_{j+1}(z) Q^o(dz|x, u) + \frac{R}{2} \left( \sup_{z \in \mathcal{X}} V_{j+1}(z) - \inf_{z \in \mathcal{X}} V_{j+1}(z) \right) \right\}.$$

Define  $\bar{V}_j(x) = V_{n-j}(x)$ , ( $j = 0, 1, \dots, n$ ). Then (4.9) may be written in the “forward” form

(4.10)

$$\bar{V}_{j+1}(x) = \inf_{u \in \mathcal{U}(x)} \left\{ f(x, u) + \int_{\mathcal{X}} \bar{V}_j(z) Q^o(dz|x, u) + \frac{R}{2} \left( \sup_{z \in \mathcal{X}} \bar{V}_j(z) - \inf_{z \in \mathcal{X}} \bar{V}_j(z) \right) \right\}.$$

Substituting (4.10) to (4.2)-(4.3), we have

$$(4.11) \quad J_{n+1}^*(x, h) = \inf_{u \in \mathcal{U}(x)} \left\{ f(x, u) + \int_{\mathcal{X}} J_n^*(z, h) Q^o(dz|x, u) + \frac{R}{2} \left( \sup_{z \in \mathcal{X}} J_n^*(z, h) - \inf_{z \in \mathcal{X}} J_n^*(z, h) \right) \right\}.$$

Thus, from (4.4) we have

$$(4.12) \quad (n+1)\rho(x) + h(x) = \inf_{u \in \mathcal{U}(x)} \left\{ f(x, u) + \int_{\mathcal{X}} (n\rho(z) + h(z)) Q^0(dz|x, u) + \frac{R}{2} \left( \sup_{z \in \mathcal{X}} (n\rho(z) + h(z)) - \inf_{z \in \mathcal{X}} (n\rho(z) + h(z)) \right) \right\}.$$

Evaluating (4.12) at  $n = 0$  we obtain (b). Furthermore, since  $\rho(\cdot)$ ,  $h(\cdot)$  and  $f(\cdot, \cdot)$  are bounded, then multiplying both sides of (4.12) by  $1/n$  and letting  $n \rightarrow \infty$  yields (a).

Finally, for any deterministic stationary policy  $g^\infty \in G_{SM}$ , we have that

$$(4.13) \quad J_{n+1}(g^\infty, x, h) = f(x, \varphi) + \int_{\mathcal{X}} J_n(g^\infty, z, h) Q^o(dz|x, \varphi) + \frac{R}{2} \left( \sup_{z \in \mathcal{X}} J_n(g^\infty, z, h) - \inf_{z \in \mathcal{X}} J_n(g^\infty, z, h) \right), \quad x \in \mathcal{X}.$$

Thus, if  $\varphi \in \mathbb{F}$  satisfies (4.4), then by (4.11)-(4.13) we have that

$$(n+1)\rho(x) + h(x) = f(x, \varphi) + \int_{\mathcal{X}} (n\rho(z) + h(z)) Q^o(dz|x, \varphi) + \frac{R}{2} \left( \sup_{z \in \mathcal{X}} (n\rho(z) + h(z)) - \inf_{z \in \mathcal{X}} (n\rho(z) + h(z)) \right)$$

which, as before, gives (4.5) and (4.6).

(Sufficiency). Conversely, suppose  $(\rho, h, \varphi)$  satisfy (a)-(c). Proceeding by induction equation (4.4) is trivially satisfied when  $n = 0$ . Suppose that is true for some  $n \geq 0$ . Then,

the following is obtained

$$\begin{aligned}
J_{n+1}^*(x, h) &= \inf_{u \in \mathcal{U}(x)} \left\{ f(x, u) + \int_{\mathcal{X}} (n\rho(z) + h(z)) Q^o(dz|x, u) \right. \\
&\quad \left. + \frac{R}{2} \left( \sup_{z \in \mathcal{X}} (n\rho(z) + h(z)) - \inf_{z \in \mathcal{X}} (n\rho(z) + h(z)) \right) \right\} \\
&= \inf_{u \in \mathcal{U}(x)} \left\{ f(x, u) + \int_{\mathcal{X}} (n\rho(z) + h(z)) Q^*(dz|x, u) \right\} \\
&\geq \inf_{u \in \mathcal{U}(x)} \left\{ f(x, u) + \int_{\mathcal{X}} h(z) Q^*(dz|x, u) \right\} + n \inf_{u \in \mathcal{U}(x)} \left\{ \int_{\mathcal{X}} \rho(z) Q^*(dz|x, u) \right\} \\
&= \inf_{u \in \mathcal{U}(x)} \left\{ f(x, u) + \int_{\mathcal{X}} h(z) Q^o(dz|x, u) + \frac{R}{2} \left( \sup_{z \in \mathcal{X}} h(z) - \inf_{z \in \mathcal{X}} h(z) \right) \right\} \\
&\quad + n \inf_{u \in \mathcal{U}(x)} \left\{ \int_{\mathcal{X}} \rho(z) Q^o(dz|x, u) + \frac{R}{2} \left( \sup_{z \in \mathcal{X}} \rho(z) - \inf_{z \in \mathcal{X}} \rho(z) \right) \right\} \\
&= (n+1)\rho(x) + h(x).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
J_{n+1}^*(x, h) &\leq J_{n+1}(g^\infty, x, h) \\
&= f(x, \varphi) + \int_{\mathcal{X}} (n\rho(z) + h(z)) Q^o(dz|x, \varphi) \\
&\quad + \frac{R}{2} \left( \sup_{z \in \mathcal{X}} (n\rho(z) + h(z)) - \inf_{z \in \mathcal{X}} (n\rho(z) + h(z)) \right) \\
&= f(x, \varphi) + \int_{\mathcal{X}} (n\rho(z) + h(z)) Q^*(dz|x, \varphi) \\
&= f(x, \varphi) + \int_{\mathcal{X}} h(z) Q^*(dz|x, \varphi) + n \int_{\mathcal{X}} \rho(z) Q^*(dz|x, \varphi) \\
&= f(x, \varphi) + \int_{\mathcal{X}} h(z) Q^o(dz|x, \varphi) + \frac{R}{2} \left( \sup_{z \in \mathcal{X}} h(z) - \inf_{z \in \mathcal{X}} h(z) \right) \\
&\quad + n \left\{ \int_{\mathcal{X}} \rho(z) Q^o(dz|x, \varphi) + \frac{R}{2} \left( \sup_{z \in \mathcal{X}} \rho(z) - \inf_{z \in \mathcal{X}} \rho(z) \right) \right\} \\
&= (n+1)\rho(x) + h(x)
\end{aligned}$$

where the second and third equalities follow by applying (2.1). This implies,  $J_{n+1}^*(x, h) = J_{n+1}(g^\infty, x, h) = (n+1)\rho(x) + h(x)$ .  $\square$

REMARK 4.3. We note that in Definition 4.1, the condition  $(\rho, h)$  are bounded continuous and non-negative can be relaxed to continuous and non-negative. In this case, if (4.4) holds, i.e.,  $(\rho, h, \varphi)$  is a canonical triplet then (a)-(c) hold.

Due to the fact that the average cost as an optimality criterion is underselective, i.e., with limitations in distinguishing optimal policies with different costs, we introduce next a more selective criterion. For other underselective and overselective optimality criteria see [10, 11].

DEFINITION 4.4. A policy  $g^\dagger$  is said to be

(a) [9] Strong average cost optimal if

$$(4.14) \quad J(g^\dagger, x) \leq \liminf_{n \rightarrow \infty} \frac{1}{n} J_n(g, x), \quad \forall g \in G, x \in \mathcal{X}.$$

(b) [11] F-strong average cost optimal if

$$(4.15) \quad \lim_{n \rightarrow \infty} \frac{1}{n} \left( J_n(g^\dagger, x) - J_n^*(x) \right) = 0, \quad \forall x \in \mathcal{X}$$

where  $J_n^*(x) = \inf_{g \in G} J_n(g, x)$ .

Based on Definition 4.4, next we derive stronger results.

THEOREM 4.5. [12] Suppose the cost function  $f$  satisfies Assumption 1.4, and let  $(\rho, h, \varphi)$  be a canonical triplet (with  $h$  not necessarily bounded).

(a) If for every  $g \in G$  and  $x \in \mathcal{X}$

$$(4.16) \quad \lim_{n \rightarrow \infty} \mathbb{E}_x^{g, Q^*} \left\{ \frac{h(x_n)}{n} \right\} = 0$$

then  $g^\infty$  is an average cost optimal policy and  $\rho$  is the average cost value function

$$(4.17) \quad J^*(x) = \rho(x) = J(g^\infty, x) = \lim_{n \rightarrow \infty} \frac{1}{n} J_n(g^\infty, x), \quad \forall x.$$

(b) If for every  $x \in \mathcal{X}$

$$(4.18) \quad \lim_{n \rightarrow \infty} \sup_{g \in G} \mathbb{E}_x^{g, Q^*} \left\{ \frac{h(x_n)}{n} \right\} = 0$$

then  $g^\infty$  is strong average cost optimal and  $F$ -strong average cost optimal and

$$(4.19) \quad J^*(x) = \lim_{n \rightarrow \infty} \frac{1}{n} J_n^*(x).$$

Proof. (a) From (4.2)-(4.3) and the last equality in (4.4)

$$n\rho(x) + h(x) = J_n^*(x, h) \leq J_n(g, x) + \mathbb{E}_x^{g, Q^*} \{h(x_n)\}, \quad \forall g \in G, x \in \mathcal{X}.$$

Hence, multiplying by  $1/n$ , taking the  $\limsup$  as  $n \rightarrow \infty$ , by virtue of (4.16), we have  $\rho(x) \leq J(g, x)$ ,  $\forall g, x$  which implies

$$(4.20) \quad \rho(x) \leq J^*(x), \quad \forall x.$$

Furthermore, from (4.4) again

$$(4.21) \quad J_n(g^\infty, x, h) = J_n(g^\infty, x) + \mathbb{E}_x^{g^\infty, Q^*} \{h(x_n)\} = n\rho(x) + h(x).$$

Finally, multiplying both sides of (4.21) by  $1/n$  and then taking both  $\limsup$  and  $\liminf$  as  $n \rightarrow \infty$ , we obtain the last two equalities in (4.17), which in turn, together with (4.20), yield the first one since  $J^*(x) \leq J(g^\infty, x)$ .

(b) The first equality in (4.4) gives

$$(4.22) \quad J_n^*(x, h) = J_n(g^\infty, x) + \mathbb{E}_x^{g^\infty, Q^*} \{h(x_n)\}.$$

On the other hand, by (4.2)-(4.3)

$$J_n^*(x, h) = \inf_{g \in G} \left( J_n(g, x) + \mathbb{E}_x^{g, Q^*} \{h(x_n)\} \right) \leq J_n^*(x) + \sup_{g \in G} \mathbb{E}_x^{g, Q^*} \{h(x_n)\}.$$

Thus,

$$(4.23) \quad 0 \leq J_n(g^\infty, x) - J_n^*(x) \leq \sup_{g \in G} \mathbb{E}_x^{g, Q^*} \{h(x_n)\} - \mathbb{E}_x^{g^\infty, Q^*} \{h(x_n)\}.$$

Hence, if  $h$  satisfies (4.18), then  $g^\infty$  is F-strong average cost optimal. Finally, to prove that  $g^\infty$  is strong average cost optimal, we use (4.22) again to obtain

$$J_n(g^\infty, x) + \mathbb{E}_x^{g^\infty, Q^*} \{h(x_n)\} \leq J_n(g, x) + \mathbb{E}_x^{g, Q^*} \{h(x_n)\}, \quad \forall g, x, n$$

so that from (4.18)

$$(4.24) \quad \liminf_{n \rightarrow \infty} \frac{1}{n} J_n(g^\infty, x) \leq \liminf_{n \rightarrow \infty} \frac{1}{n} J_n(g, x).$$

Since the left-hand side equals to  $J(g^\infty, x)$  (see (4.17)) it follows that  $g^\infty$  is indeed strong average cost optimal and the proof is complete.  $\square$

Note that, in the case in which  $\rho(\cdot)$  is constant, that is  $\rho$  does not vary with  $x$ , then the first optimality equation of Theorem 4.2 is redundant and hence (a)-(c) reduce to

$$(4.25) \quad \rho^* + h(x) = \inf_{u \in \mathcal{U}(x)} \left\{ f(x, u) + \int_{\mathcal{X}} h(z) Q^o(dz|x, u) + \frac{R}{2} \left( \sup_{z \in \mathcal{X}} h(z) - \inf_{z \in \mathcal{X}} h(z) \right) \right\}$$

$$(4.26) \quad \rho^* + h(x) = f(x, \varphi) + \int_{\mathcal{X}} h(z) Q^o(dz|x, \varphi) + \frac{R}{2} \left( \sup_{z \in \mathcal{X}} h(z) - \inf_{z \in \mathcal{X}} h(z) \right).$$

Next, we use equations (a')-(c') of Theorem 4.2 to develop a general policy iteration algorithm for average cost dynamic programming.

**4.2. General Policy Iteration Algorithm for Finite Alphabet Spaces.** In this section, we provide a policy iteration algorithm to obtain average cost optimal policies, in which policy evaluation and policy improvement steps are evaluated using the maximizing conditional distribution given by (3.26). The proposed algorithm is considerably more complex compared to Algorithm 3.9. Nevertheless, it solves the MDP for all range of values of total variation parameter  $R \in [0, 2]$ , and without imposing the irreducibility condition, as in Section 3.1.1.

ALGORITHM 4.6. (*General policy iteration*)

- 1) Let  $m = 0$  and select an arbitrary stationary Markov control policy  $g_0 : \mathcal{X} \mapsto \mathcal{U}$ .
- 2) (*Policy Evaluation*) Solve the equations

$$(4.27) \quad J_{Q^o}(g_m) = Q^o(g_m) J_{Q^o}(g_m)$$

$$(4.28) \quad J_{Q^o}(g_m) + h_{Q^o}(g_m) = f(g_m) + Q^o(g_m) h_{Q^o}(g_m)$$

for  $J_{Q^o}(g_m)$  and  $h_{Q^o}(g_m)$ . Identify the support sets of (4.28) using (3.27) (where  $h$  replaces  $V$ ), and construct the matrix  $Q^*(g_m)$  using (3.26). Solve the equations

$$(4.29) \quad J_{Q^*}(g_m) = Q^*(g_m) J_{Q^*}(g_m)$$

$$(4.30) \quad J_{Q^*}(g_m) + h_{Q^*}(g_m) = f(g_m) + Q^*(g_m) h_{Q^*}(g_m)$$

for  $J_{Q^*}(g_m)$  and  $h_{Q^*}(g_m)$ .

- 3) (*Policy Improvement*)

a) Let

$$(4.31) \quad g_{m+1} = \arg \min_{g \in \mathbb{R}^{|\mathcal{X}|}} \{Q^*(g) J_{Q^*}(g_m)\}.$$

If  $g_{m+1} = g_m$  go to step 3b); otherwise let  $m = m + 1$  and return to step 2.

b) Let

$$(4.32) \quad g_{m+1} = \arg \min_{g \in \mathbb{R}^{|\mathcal{X}|}} \{f(g) + Q^*(g) h_{Q^*}(g_m)\}.$$

4) If  $g_{m+1} = g_m$ , let  $g^* = g_m$ ; else let  $m = m + 1$  and return to step 2.

For MCP with finite state and action spaces the proposed general policy iteration algorithm converges in a finite number of iterations. However, for MCP on Borel spaces the proposed policy iteration algorithm might not converge, or it might converge to a suboptimal value, and hence one must introduce additional assumptions (i.e., see [13, 15]). In Section 5.2, we illustrate through an example how Algorithm 4.6 is applied.

**5. Examples.** In this section we illustrate the new dynamic programming equations and the corresponding policy iteration algorithms through examples. In particular, in Section 5.1 we present an application of the infinite horizon minimax problem for average cost by employing policy iteration Algorithm 3.9, and in Section 5.2 we present an application of the infinite horizon minimax problem for average cost by employing policy iteration Algorithm 4.6. The essential difference between the two examples is that the MDP of the latter is described by a transition probability graph which is reducible.

**5.1. Infinite Horizon Minimax MDP - Policy Iteration Algorithm 3.9.** Here, we illustrate an application of the infinite horizon minimax problem for average cost, by considering the stochastic control system as shown in Fig.3.1, with state space  $\mathcal{X} = \{1, 2, 3\}$  and control set  $\mathcal{U} = \{u_1, u_2\}$ . Assume that the nominal transition probabilities under controls  $u_1$  and  $u_2$  are given by

$$(5.1) \quad Q^o(u_1) = \frac{1}{9} \begin{pmatrix} 3 & 1 & 5 \\ 4 & 2 & 3 \\ 1 & 6 & 2 \end{pmatrix}, \quad Q^o(u_2) = \frac{1}{9} \begin{pmatrix} 1 & 2 & 6 \\ 4 & 2 & 3 \\ 4 & 1 & 4 \end{pmatrix}$$

the total variation distance radius is  $R = 6/9$ , and the cost function under each state and action is

$$f(1, u_1) = 2, f(2, u_1) = 1, f(3, u_1) = 3, f(1, u_2) = 0.5, f(2, u_2) = 3, f(3, u_2) = 0.$$

To obtain an optimal stationary policy of the infinite horizon minimax problem for average cost, policy iteration algorithm 3.9 is applied.

**A. Let  $m = 0$ .**

1) Select the initial policies as follows  $g_0(1) = u_1, g_0(2) = u_2, g_0(3) = u_2$ .

2) Solve the equation  $J_{Q^o}(g_0)e + V_{Q^o}(g_0) = f(g_0) + Q^o(g_0)V_{Q^o}(g_0)$  for  $J_{Q^o}(g_0) \in \mathbb{R}$  and  $V_{Q^o}(g_0) \in \mathbb{R}^3$ , which is given by

$$J_{Q^o}(g_0) \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} V_{Q^o}(g_0, 1) \\ V_{Q^o}(g_0, 2) \\ V_{Q^o}(g_0, 3) \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix} + \frac{1}{9} \begin{pmatrix} 3 & 1 & 5 \\ 4 & 2 & 3 \\ 4 & 1 & 4 \end{pmatrix} \begin{pmatrix} V_{Q^o}(g_0, 1) \\ V_{Q^o}(g_0, 2) \\ V_{Q^o}(g_0, 3) \end{pmatrix}.$$

Since  $V_{Q^o}(g_0)$  is uniquely determined up to an additive constant, let  $V_{Q^o}(g_0, 3) = 0$ . The solution is

$$\begin{pmatrix} V_{Q^o}(g_0, 1) \\ V_{Q^o}(g_0, 2) \\ V_{Q^o}(g_0, 3) \end{pmatrix} = \begin{pmatrix} 1.8 \\ 3.375 \\ 0 \end{pmatrix}, \quad J_{Q^o}(g_0) = 1.175.$$

Note that,  $V_{Q^\circ} \triangleq \{V_{Q^\circ}(1), V_{Q^\circ}(2), V_{Q^\circ}(3)\}$ ,  $|\mathcal{X}| = 3$ , and hence

$$\begin{aligned}\mathcal{X}^+ &\triangleq \{x \in \mathcal{X} : V_{Q^\circ}(x) = \max\{V_{Q^\circ}(x) : x \in \mathcal{X}\}\} \\ &= \{x \in \mathcal{X} : V_{Q^\circ}(x) = V_{Q^\circ}(2)\} = \{2\} \\ \mathcal{X}^- &\triangleq \{x \in \mathcal{X} : V_{Q^\circ}(x) = \min\{V_{Q^\circ}(x) : x \in \mathcal{X}\}\} \\ &= \{x \in \mathcal{X} : V_{Q^\circ}(x) = V_{Q^\circ}(3)\} = \{3\} \\ \mathcal{X}_1 &\triangleq \{x \in \mathcal{X} : V_{Q^\circ}(x) = \min\{V_{Q^\circ}(\alpha) : \alpha \in \mathcal{X} \setminus \mathcal{X}^+ \cup \mathcal{X}^-\}\} \\ &= \{x \in \mathcal{X} : V_{Q^\circ}(x) = V_{Q^\circ}(1)\} = \{1\}.\end{aligned}$$

Once the partition is been identified, (3.26) is applied to obtain (5.2) and (5.3).

$$\begin{aligned}(5.2) \quad Q^*(u_1) &= \begin{pmatrix} \left( q_{11}^\circ(u_1) - \left( \frac{R}{2} - q_{13}^\circ(u_1) \right)^+ \right)^+ & \min \left( 1, q_{12}^\circ(u_1) + \frac{R}{2} \right) & \left( q_{13}^\circ(u_1) - \frac{R}{2} \right)^+ \\ \left( q_{21}^\circ(u_1) - \left( \frac{R}{2} - q_{23}^\circ(u_1) \right)^+ \right)^+ & \min \left( 1, q_{22}^\circ(u_1) + \frac{R}{2} \right) & \left( q_{23}^\circ(u_1) - \frac{R}{2} \right)^+ \\ \left( q_{31}^\circ(u_1) - \left( \frac{R}{2} - q_{33}^\circ(u_1) \right)^+ \right)^+ & \min \left( 1, q_{32}^\circ(u_1) + \frac{R}{2} \right) & \left( q_{33}^\circ(u_1) - \frac{R}{2} \right)^+ \end{pmatrix} \\ &= \frac{1}{9} \begin{pmatrix} 3 & 4 & 2 \\ 4 & 5 & 0 \\ 0 & 9 & 0 \end{pmatrix}.\end{aligned}$$

$$\begin{aligned}(5.3) \quad Q^*(u_2) &= \begin{pmatrix} \left( q_{11}^\circ(u_2) - \left( \frac{R}{2} - q_{13}^\circ(u_2) \right)^+ \right)^+ & \min \left( 1, q_{12}^\circ(u_2) + \frac{R}{2} \right) & \left( q_{13}^\circ(u_2) - \frac{R}{2} \right)^+ \\ \left( q_{21}^\circ(u_2) - \left( \frac{R}{2} - q_{23}^\circ(u_2) \right)^+ \right)^+ & \min \left( 1, q_{22}^\circ(u_2) + \frac{R}{2} \right) & \left( q_{23}^\circ(u_2) - \frac{R}{2} \right)^+ \\ \left( q_{31}^\circ(u_2) - \left( \frac{R}{2} - q_{33}^\circ(u_2) \right)^+ \right)^+ & \min \left( 1, q_{32}^\circ(u_2) + \frac{R}{2} \right) & \left( q_{33}^\circ(u_2) - \frac{R}{2} \right)^+ \end{pmatrix} \\ &= \frac{1}{9} \begin{pmatrix} 1 & 5 & 3 \\ 4 & 5 & 0 \\ 4 & 4 & 1 \end{pmatrix}.\end{aligned}$$

The transition probability graph of  $Q^*$ , under controls  $u_1$  and  $u_2$ , is depicted in Fig 5.1. Note that, since every state can reach every other state, matrix  $Q^*(u)$  remains irreducible under both controls. Next, we proceed to solve the equation  $J_{Q^*}(g_0)e + V_{Q^*}(g_0) = f(g_0) +$

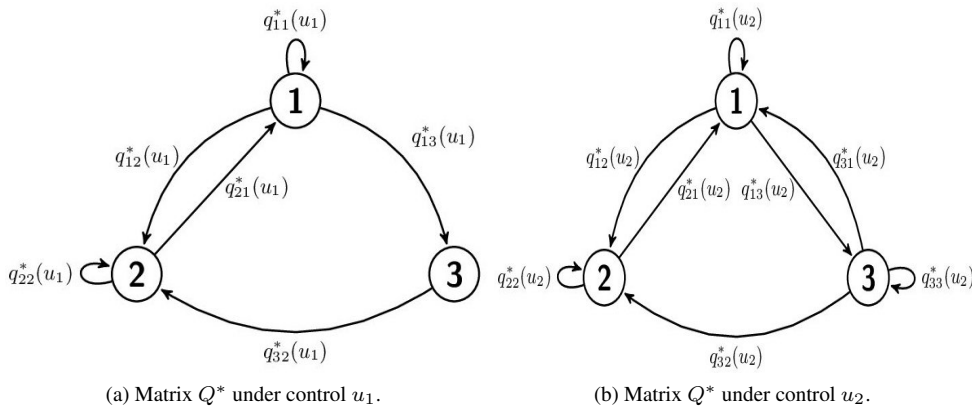


FIG. 5.1. Transition Probability Graph of  $Q^*$  under controls  $u_1$  and  $u_2$ .



$Q^*(g_0)V_{Q^*}(g_0)$  for  $J_{Q^*}(g_0) \in \mathbb{R}$  and  $V_{Q^*}(g_0) \in \mathbb{R}^3$ , which is given by

$$J_{Q^*}(g_0) \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} V_{Q^*}(g_0, 1) \\ V_{Q^*}(g_0, 2) \\ V_{Q^*}(g_0, 3) \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix} + \frac{1}{9} \begin{pmatrix} 3 & 4 & 2 \\ 4 & 5 & 0 \\ 4 & 4 & 1 \end{pmatrix} \begin{pmatrix} V_{Q^*}(g_0, 1) \\ V_{Q^*}(g_0, 2) \\ V_{Q^*}(g_0, 3) \end{pmatrix}.$$

Since  $V_{Q^*}(g_0)$  is uniquely determined up to an additive constant, let  $V_{Q^*}(g_0, 3) = 0$ . The solution is

$$\begin{pmatrix} V_{Q^*}(g_0, 1) \\ V_{Q^*}(g_0, 2) \\ V_{Q^*}(g_0, 3) \end{pmatrix} = \begin{pmatrix} 1.8 \\ 3.375 \\ 0 \end{pmatrix}, \quad J_{Q^*}(g_0) = 2.3.$$

3) Let  $g_1 = \operatorname{argmin}_{g \in \mathbb{R}^3} \{f(g) + Q^*(g)V_{Q^*}(g_0)\}$ . Then

$$\begin{aligned} g_1(1) &= \operatorname{argmin} \left\{ f(1, u_1) + q_{11}^*(u_1)V_{Q^*}(g_0, 1) + q_{12}^*(u_1)V_{Q^*}(g_0, 2) + q_{13}^*(u_1)V_{Q^*}(g_0, 3), \right. \\ &\quad \left. f(1, u_2) + q_{11}^*(u_2)V_{Q^*}(g_0, 1) + q_{12}^*(u_2)V_{Q^*}(g_0, 2) + q_{13}^*(u_2)V_{Q^*}(g_0, 3) \right\} \\ &= \operatorname{argmin} \{4.099, 2.573\} = \{2\} \implies g_1(1) = u_2. \end{aligned}$$

Following a similar procedure for the rest we obtain the following.

$$g_1(2) = \operatorname{argmin} \{3.673, 5.673\} = \{1\} \implies g_1(2) = u_1$$

$$g_1(3) = \operatorname{argmin} \{6.375, 2.3\} = \{2\} \implies g_1(3) = u_2.$$

Since,  $g_1 \neq g_0$ , let  $m = 1$  and return to step 2.

**B. Let  $m = 1$ .**

2) Solve the equation  $J_{Q^o}(g_1)e + V_{Q^o}(g_1) = f(g_1) + Q^o(g_1)V_{Q^o}(g_1)$ ,  $V_{Q^o}(g_1, 3) = 0$ , for  $J_{Q^o}(g_1) \in \mathbb{R}$  and  $V_{Q^o}(g_1) \in \mathbb{R}^3$ . The solution is

$$\begin{pmatrix} V_{Q^o}(g_1, 1) \\ V_{Q^o}(g_1, 2) \\ V_{Q^o}(g_1, 3) \end{pmatrix} = \begin{pmatrix} 0.468 \\ 1.125 \\ 0 \end{pmatrix}, \quad J_{Q^o}(g_1) = 0.333.$$

Therefore,  $\mathcal{X}^+ = \{2\}$ ,  $\mathcal{X}^- = \{3\}$  and  $\mathcal{X}_1 = \{1\}$ . Since the partition is the same as in  $m = 0$  then  $Q^*(u_1)$  and  $Q^*(u_2)$  are given by (5.2) and (5.3), respectively.

Solve the equation  $J_{Q^*}(g_1)e + V_{Q^*}(g_1) = f(g_1) + Q^*(g_1)V_{Q^*}(g_1)$ ,  $V_{Q^*}(g_1, 3) = 0$ , for  $J_{Q^*}(g_1) \in \mathbb{R}$  and  $V_{Q^*}(g_1) \in \mathbb{R}^3$ . The solution is

$$\begin{pmatrix} V_{Q^*}(g_1, 1) \\ V_{Q^*}(g_1, 2) \\ V_{Q^*}(g_1, 3) \end{pmatrix} = \begin{pmatrix} 0.468 \\ 1.125 \\ 0 \end{pmatrix}, \quad J_{Q^*}(g_1) = 0.708.$$

3) Let  $g_2 = \operatorname{argmin}_{g \in \mathbb{R}^3} \{f(g) + Q^*(g)V_{Q^*}(g_1)\}$ . Then

$$g_2(1) = \operatorname{argmin} \{2.656, 1.177\} = \{2\} \implies g_2(1) = u_2$$

$$g_2(2) = \operatorname{argmin} \{1.831, 3.831\} = \{1\} \implies g_2(2) = u_1$$

$$g_2(3) = \operatorname{argmin} \{4.125, 0.708\} = \{2\} \implies g_2(3) = u_2.$$

4) Since,  $g_2 = g_1$ , then  $g^* = g_1$  is an optimal control policy with  $J_{Q^*} = 0.708$ ,  $V_{Q^*}(1) = 0.468$ ,  $V_{Q^*}(2) = 1.125$  and  $V_{Q^*}(3) = 0$ .

**5.2. Infinite Horizon Minimax MDP - General Policy Iteration Algorithm 4.6.** In this example, we illustrate an application of the infinite horizon minimax problem for average cost, by considering the stochastic control system shown in Fig.3.1, with  $\mathcal{X} = \{1, 2, 3\}$  and control set  $\mathcal{U} = \{u_1, u_2\}$ . The essential difference between this example and the previous one, is that here, the stochastic control system under consideration is described by a transition probability graph which is reducible, and hence general policy iteration algorithm 4.6 is applied.

Assume that the nominal transition probabilities under controls  $u_1$  and  $u_2$  are given by

$$(5.4) \quad Q^o(u_1) = \frac{1}{9} \begin{pmatrix} 0 & 5 & 4 \\ 0 & 9 & 0 \\ 0 & 0 & 9 \end{pmatrix}, \quad Q^o(u_2) = \frac{1}{9} \begin{pmatrix} 2 & 7 & 0 \\ 3 & 6 & 0 \\ 8 & 0 & 1 \end{pmatrix}$$

the total variation distance radius is  $R = 14/9$ , and the cost function under each state and action is

$$f(1, u_1) = 2, \quad f(2, u_1) = 1, \quad f(3, u_1) = 3, \quad f(1, u_2) = 0.5, \quad f(2, u_2) = 3, \quad f(3, u_2) = 0.$$

**A. Let  $m = 0$ .**

1) Select the initial policies as follows  $g_0(1) = u_1, g_0(2) = u_1, g_0(3) = u_1$ .

2) Solve the equation  $J_{Q^o}(g_0) = Q^o(g_0)J_{Q^o}(g_0)$ . The optimality equations (4.27) are

$$(5.5a) \quad J_{Q^o}(g_0, 1) = \frac{5}{9}J_{Q^o}(g_0, 2) + \frac{4}{9}J_{Q^o}(g_0, 3)$$

$$(5.5b) \quad J_{Q^o}(g_0, 2) = J_{Q^o}(g_0, 2)$$

$$(5.5c) \quad J_{Q^o}(g_0, 3) = J_{Q^o}(g_0, 3).$$

Next, solve the equation  $J_{Q^o}(g_0) + h_{Q^o}(g_0) = f(g_0) + Q^o(g_0)h_{Q^o}(g_0)$ , for  $J_{Q^o}(g_0) \in \mathbb{R}^3$  and  $h_{Q^o}(g_0) \in \mathbb{R}^3$ . The optimality equations (4.28) are given by

$$(5.6a) \quad J_{Q^o}(g_0, 1) + h_{Q^o}(g_0, 1) = 2 + \frac{5}{9}h_{Q^o}(g_0, 2) + \frac{4}{9}h_{Q^o}(g_0, 3)$$

$$(5.6b) \quad J_{Q^o}(g_0, 2) + h_{Q^o}(g_0, 2) = 1 + h_{Q^o}(g_0, 2)$$

$$(5.6c) \quad J_{Q^o}(g_0, 3) + h_{Q^o}(g_0, 3) = 3 + h_{Q^o}(g_0, 3).$$

The solution of (5.5) and (5.6) has

$$\begin{aligned} h_{Q^o}(g_0, 1) &= \frac{1}{9} + \frac{5}{9}\alpha + \frac{4}{9}\beta, & h_{Q^o}(g_0, 2) &= \alpha, & h_{Q^o}(g_0, 3) &= \beta, \\ J_{Q^o}(g_0, 1) &= 1.888, & J_{Q^o}(g_0, 2) &= 1, & J_{Q^o}(g_0, 3) &= 3. \end{aligned}$$

Setting  $\alpha = 1$  and  $\beta = 0$  (arbitrary constants) yields

$$h_{Q^o}(g_0, 1) = 0.666, \quad h_{Q^o}(g_0, 2) = 1, \quad h_{Q^o}(g_0, 3) = 0.$$

Note that,  $h_{Q^o} = \{h_{Q^o}(1), h_{Q^o}(2), h_{Q^o}(3)\}$ , and hence the support sets based on the values of  $h_{Q^o}$  are  $\mathcal{X}^+ = \{2\}$ ,  $\mathcal{X}^- = \{3\}$  and  $\mathcal{X}_1 = \{1\}$ . Once the partition is been identified, (3.26)

is applied to obtain (5.7) and (5.8).

$$(5.7) \quad \begin{aligned} Q^*(u_1) &= \begin{pmatrix} \left( q_{11}^o(u_1) - \left( \frac{R}{2} - q_{13}^o(u_1) \right)^+ \right)^+ & \min \left( 1, q_{12}^o(u_1) + \frac{R}{2} \right) & \left( q_{13}^o(u_1) - \frac{R}{2} \right)^+ \\ \left( q_{21}^o(u_1) - \left( \frac{R}{2} - q_{23}^o(u_1) \right)^+ \right)^+ & \min \left( 1, q_{22}^o(u_1) + \frac{R}{2} \right) & \left( q_{23}^o(u_1) - \frac{R}{2} \right)^+ \\ \left( q_{31}^o(u_1) - \left( \frac{R}{2} - q_{33}^o(u_1) \right)^+ \right)^+ & \min \left( 1, q_{32}^o(u_1) + \frac{R}{2} \right) & \left( q_{33}^o(u_1) - \frac{R}{2} \right)^+ \end{pmatrix} \\ &= \frac{1}{9} \begin{pmatrix} 0 & 9 & 0 \\ 0 & 9 & 0 \\ 0 & 7 & 2 \end{pmatrix} \end{aligned}$$

$$(5.8) \quad \begin{aligned} Q^*(u_2) &= \begin{pmatrix} \left( q_{11}^o(u_2) - \left( \frac{R}{2} - q_{13}^o(u_2) \right)^+ \right)^+ & \min \left( 1, q_{12}^o(u_2) + \frac{R}{2} \right) & \left( q_{13}^o(u_2) - \frac{R}{2} \right)^+ \\ \left( q_{21}^o(u_2) - \left( \frac{R}{2} - q_{23}^o(u_2) \right)^+ \right)^+ & \min \left( 1, q_{22}^o(u_2) + \frac{R}{2} \right) & \left( q_{23}^o(u_2) - \frac{R}{2} \right)^+ \\ \left( q_{31}^o(u_2) - \left( \frac{R}{2} - q_{33}^o(u_2) \right)^+ \right)^+ & \min \left( 1, q_{32}^o(u_2) + \frac{R}{2} \right) & \left( q_{33}^o(u_2) - \frac{R}{2} \right)^+ \end{pmatrix} \\ &= \frac{1}{9} \begin{pmatrix} 0 & 9 & 0 \\ 0 & 9 & 0 \\ 2 & 7 & 0 \end{pmatrix}. \end{aligned}$$

Next, solve the equation  $J_{Q^*}(g_0) = Q^*(g_0)J_{Q^*}(g_0)$ . The optimality equations (4.29) are

$$(5.9a) \quad J_{Q^*}(g_0, 1) = J_{Q^*}(g_0, 2)$$

$$(5.9b) \quad J_{Q^*}(g_0, 2) = J_{Q^*}(g_0, 3)$$

$$(5.9c) \quad J_{Q^*}(g_0, 3) = \frac{7}{9}J_{Q^*}(g_0, 2) + \frac{2}{9}J_{Q^*}(g_0, 3)$$

and hence,  $J_{Q^*}(g_0, 1) = J_{Q^*}(g_0, 2) = J_{Q^*}(g_0, 3)$ .

Next, solve the equation  $J_{Q^*}(g_0) + h_{Q^*}(g_0) = f(g_0) + Q^*(g_0)h_{Q^*}(g_0)$ , for  $J_{Q^*}(g_0) \in \mathbb{R}^3$  and  $h_{Q^*}(g_0) \in \mathbb{R}^3$ . The optimality equations (4.30) are given by

$$(5.10a) \quad J_{Q^*}(g_0, 1) + h_{Q^*}(g_0, 1) = 2 + h_{Q^*}(g_0, 2)$$

$$(5.10b) \quad J_{Q^*}(g_0, 2) + h_{Q^*}(g_0, 2) = 1 + h_{Q^*}(g_0, 3)$$

$$(5.10c) \quad J_{Q^*}(g_0, 3) + \frac{7}{9}h_{Q^*}(g_0, 3) = 3 + \frac{2}{9}h_{Q^*}(g_0, 2).$$

The solution of (5.9) and (5.10) has

$$\begin{aligned} h_{Q^*}(g_0, 1) &= 1 + \alpha, & h_{Q^*}(g_0, 2) &= \alpha, & h_{Q^*}(g_0, 3) &= \frac{18}{7} + \alpha, \\ J_{Q^*}(g_0, 1) &= 1, & J_{Q^*}(g_0, 2) &= 1, & J_{Q^*}(g_0, 3) &= 1. \end{aligned}$$

Setting  $\alpha = 1$  (arbitrary constant) yields

$$h_{Q^*}(g_0, 1) = 2, \quad h_{Q^*}(g_0, 2) = 1, \quad h_{Q^*}(g_0, 3) = 3.57.$$

3) a) Since  $J_{Q^*}(g_0, 1) = J_{Q^*}(g_0, 2) = J_{Q^*}(g_0, 3)$ , then clearly  $g_1 = g_0$  and we proceed to step 3b).

b) Let  $g_1 = \operatorname{argmin}_{g \in \mathbb{R}^3} \{f(g) + Q^*(g)h_{Q^*}(g_0)\}$ , then the resulting control policies are  $g_1(1) = u_2$ ,  $g_1(2) = u_1$  and  $g_1(3) = u_2$ . Since  $g_1 \neq g_0$ , let  $m = 1$  and return to step 2.

**B. Let  $m = 1$ .**

2) Solve the equation  $J_{Q^\circ}(g_1) = Q^\circ(g_1)J_{Q^\circ}(g_1)$ . The optimality equations (4.27) are

$$(5.11a) \quad J_{Q^\circ}(g_1, 1) = \frac{2}{9}J_{Q^\circ}(g_1, 1) + \frac{7}{9}J_{Q^\circ}(g_1, 2)$$

$$(5.11b) \quad J_{Q^\circ}(g_1, 2) = J_{Q^\circ}(g_1, 2)$$

$$(5.11c) \quad J_{Q^\circ}(g_1, 3) = \frac{8}{9}J_{Q^\circ}(g_1, 1) + \frac{1}{9}J_{Q^\circ}(g_1, 3)$$

and hence,  $J_{Q^\circ}(g_1, 1) = J_{Q^\circ}(g_1, 2) = J_{Q^\circ}(g_1, 3)$ .

Next, solve the equation  $J_{Q^\circ}(g_1) + h_{Q^\circ}(g_1) = f(g_1) + Q^\circ(g_1)h_{Q^\circ}(g_1)$ , for  $J_{Q^\circ}(g_1) \in \mathbb{R}^3$  and  $h_{Q^\circ}(g_1) \in \mathbb{R}^3$ . The optimality equations (4.28) are given by

$$(5.12a) \quad J_{Q^\circ}(g_1, 1) + \frac{7}{9}h_{Q^\circ}(g_1, 1) = 0.5 + \frac{7}{9}h_{Q^\circ}(g_1, 2)$$

$$(5.12b) \quad J_{Q^\circ}(g_1, 2) + h_{Q^\circ}(g_1, 2) = 1 + h_{Q^\circ}(g_1, 2)$$

$$(5.12c) \quad J_{Q^\circ}(g_1, 3) + \frac{8}{9}h_{Q^\circ}(g_1, 3) = \frac{8}{9}h_{Q^\circ}(g_1, 1).$$

The solution of (5.11) and (5.12) has

$$\begin{aligned} h_{Q^\circ}(g_1, 1) &= \alpha + \frac{9}{8}, & h_{Q^\circ}(g_1, 2) &= \alpha + \frac{99}{56}, & h_{Q^\circ}(g_1, 3) &= \alpha, \\ J_{Q^\circ}(g_1, 1) &= 1, & J_{Q^\circ}(g_1, 2) &= 1, & J_{Q^\circ}(g_1, 3) &= 1. \end{aligned}$$

Setting  $\alpha = 1$  (arbitrary constant) yields

$$h_{Q^\circ}(g_1, 1) = 2.125, \quad h_{Q^\circ}(g_1, 2) = 2.76, \quad h_{Q^\circ}(g_1, 3) = 1.$$

Hence, we proceed with the identification of the support sets, which are  $\mathcal{X}^+ = \{2\}$ ,  $\mathcal{X}^- = \{3\}$  and  $\mathcal{X}_1 = \{1\}$ . Since the partition is the same as in  $m = 0$  then  $Q^*(u_1)$  and  $Q^*(u_2)$  are equal to (5.7) and (5.8), respectively.

Next, solve the equation  $J_{Q^*}(g_1) = Q^*(g_1)J_{Q^*}(g_1)$ . The optimality equations (4.29) are

$$(5.13a) \quad J_{Q^*}(g_1, 1) = J_{Q^*}(g_1, 2)$$

$$(5.13b) \quad J_{Q^*}(g_1, 2) = J_{Q^*}(g_1, 2)$$

$$(5.13c) \quad J_{Q^*}(g_1, 3) = \frac{2}{9}J_{Q^*}(g_1, 1) + \frac{7}{9}J_{Q^*}(g_1, 2)$$

and hence,  $J_{Q^*}(g_1, 1) = J_{Q^*}(g_1, 2) = J_{Q^*}(g_1, 3)$ .

Next, solve the equation  $J_{Q^*}(g_1) + h_{Q^*}(g_1) = f(g_1) + Q^*(g_1)h_{Q^*}(g_1)$ , for  $J_{Q^*}(g_1) \in \mathbb{R}^3$  and  $h_{Q^*}(g_1) \in \mathbb{R}^3$ . The optimality equations (4.30) are given by

$$(5.14a) \quad J_{Q^*}(g_1, 1) + h_{Q^*}(g_1, 1) = 0.5 + h_{Q^*}(g_1, 2)$$

$$(5.14b) \quad J_{Q^*}(g_1, 2) + h_{Q^*}(g_1, 2) = 1 + h_{Q^*}(g_1, 2)$$

$$(5.14c) \quad J_{Q^*}(g_1, 3) + h_{Q^*}(g_1, 3) = \frac{2}{9}h_{Q^*}(g_1, 1) + \frac{7}{9}h_{Q^*}(g_1, 2).$$

The solution of (5.13) and (5.14) has

$$\begin{aligned} h_{Q^*}(g_1, 1) &= \alpha + \frac{11}{18}, & h_{Q^*}(g_1, 2) &= \alpha + \frac{10}{9}, & h_{Q^*}(g_1, 3) &= \alpha, \\ J_{Q^*}(g_1, 1) &= 1, & J_{Q^*}(g_1, 2) &= 1, & J_{Q^*}(g_1, 3) &= 1. \end{aligned}$$

Setting  $\alpha = 1$  yields

$$h_{Q^*}(g_1, 1) = 1.611, \quad h_{Q^*}(g_1, 2) = 2.111, \quad h_{Q^*}(g_1, 3) = 1.$$

3) a) Since  $J_{Q^*}(g_1, 1) = J_{Q^*}(g_1, 2) = J_{Q^*}(g_1, 3)$ , then clearly  $g_2 = g_1$  and we proceed to step 3b).

b) Let  $g_2 = \operatorname{argmin}_{g \in \mathbb{R}^3} \{f(g) + Q^*(g)h_{Q^*}(g_1)\}$ , the resulting control policies are  $g_2(1) = u_2$ ,  $g_2(2) = u_1$  and  $g_2(3) = u_2$ .

4) Because,  $g_2 = g_1$ , then  $g^* = g_1$  is an optimal control policy with  $J_{Q^*}(1) = J_{Q^*}(2) = J_{Q^*}(3) = 1$ ,  $h_{Q^*}(1) = 1.611$ ,  $h_{Q^*}(2) = 2.111$  and  $h_{Q^*}(3) = 1$ .

**6. Conclusions.** In this paper, we examined the optimality of minimax MDP via dynamic programming on an infinite horizon, when the ambiguity class is described by the total variation distance between the conditional distribution of the true controlled process and the conditional distribution of a nominal controlled process. As optimality criterion we considered the average pay-off per unit time. Under the assumption that for every stationary Markov control policy the maximizing stochastic matrix is irreducible, we derived a new dynamic programming equation and a new policy iteration algorithm. However, due to the water-filling behavior of the maximizing conditional distribution, it turns out that our proposed method of solution is limited only to a specific range of values of total variation distance. To circumvent this limit, we consider general Borel spaces, and we derive a general dynamic programming equation by introducing a pair of dynamic programming equations, and, consequently a new policy iteration algorithm, which solve the minimax MDP for all  $R \in [0, 2]$ . Finally, the application of our recommended policy iteration algorithms is shown via illustrative examples.

#### REFERENCES

- [1] A. ARAPOSTATHIS, V. S. BORKAR, E. FERNANDEZ-GAUCHERAND, M. K. GHOSH, AND S. I. MARCUS, *Discrete-time controlled Markov processes with average cost criterion: a survey*, SIAM J. Control Optim., 31 (1993), pp. 282–344.
- [2] V. S. BORKAR, *On minimum cost per unit time control of Markov chains*, SIAM J. Control Optim., 22 (1984), pp. 965–978.
- [3] ———, *Control of Markov chains with long-run average cost criterion*, Stochastic Differential systems, Stochastic Control Theory and Applications, (1988), pp. 57–77.
- [4] ———, *Control of Markov chains with long-run average cost criterion: the dynamic programming equations*, SIAM J. Control Optim., 27 (1989), pp. 642–657.
- [5] P. E. CAINES, *Linear stochastic systems*, John Wiley & Sons, Inc., New York, 1988.
- [6] C.D. CHARALAMBOUS, I. TZORTZIS, AND T. CHARALAMBOUS, *Dynamic programming with total variational distance uncertainty*, in 51st IEEE Conference on Decision and Control, Maui, Hawaii, Dec. 10–13, 2012.
- [7] CHARALAMBOS D. CHARALAMBOUS, I. TZORTZIS, S. LOYKA, AND T. CHARALAMBOUS, *Extremum problems with total variation distance and their applications*, IEEE Trans. Autom. Control, 59 (2014), pp. 2353–2368.
- [8] T.M. COVER AND J.A. THOMAS, *Elements of information theory*, John Wiley and Sons, Inc., 1991.
- [9] E. B. DYNKIN AND A. A. YUSHKEVICH, *Controlled Markov processes*, Springer-Verlag, New York, 1979.
- [10] J. FLYNN, *Conditions for the equivalence of optimality criteria in dynamic programming*, Ann. Statist., 4 (1976), pp. 936–953.
- [11] ———, *On optimality criteria for dynamic programs with long finite horizons*, J. Math. Anal. Appl., 76 (1980), pp. 202–208.
- [12] O. HERNANDEZ-LERMA AND J. B. LASSERRE, *Discrete-time Markov control processes: Basic optimality criteria*, no. v. 1 in Applications of Mathematics Stochastic Modelling and Applied Probability, Springer Verlag, 1996.
- [13] ———, *Policy iteration for average cost Markov control processes on Borel spaces*, Acta Applicandae Mathematica, 47 (1997), pp. 125–154.
- [14] P. R. KUMAR AND P. VARAIYA, *Stochastic systems: Estimation, identification, and adaptive control*, Prentice Hall, 1986.

- [15] S. P. MEYN, *The policy improvement algorithm for markov decision processes with general state space*, IEEE Trans. Autom. Control, 42 (1997), pp. 1663–1680.
- [16] M. L. PUTERMAN, *Markov decision Processes*, Wiley, New York, 1994.
- [17] M. SCHAL, *On the second optimality equation for semi-Markov decision models*, Math. Op. Res., 17 (1992), pp. 470–486.
- [18] L. I. SENNOTT, *Another set of conditions for average optimality in Markov control processes*, Systems and Control Letters, 24 (1995), pp. 147–151.
- [19] I. TZORTZIS, C. D. CHARALAMBOUS, AND T. CHARALAMBOUS, *Dynamic Programming Subject to Total Variation Distance Ambiguity*, ArXiv e-prints, (2014).
- [20] J. H. VAN SCHUPPEN, *Mathematical control and system theory of discrete-time stochastic systems*, Preprint, 2014.
- [21] A. A. YUSHKEVICH, *On a class of strategies in general Markov decision models*, Theory Probab. Appl., 18 (1973), pp. 777–779.